

# *The* **Journal of Educational Sociology**

*A Magazine of Theory and Practice*

---

## THE PENN STATE EXPERIMENTS IN CHARACTER EDUCATION

CHARLES C. PETERS, *Issue Editor*

Editorial . . . . .	213
The Potency of Instruction in Character Education <i>Charles C. Peters</i>	214
The Reliability and Validity of Estimates (Ratings) as Measuring Tools <i>James C. Swab and Charles C. Peters</i>	224
Can Social Leadership be Improved by Instruction in its Tech- nique? <i>George A. Eichler and Robert R. Merrill</i>	233
The Effect of Direct Instruction <i>E. K. Robb and J. F. Faust</i>	237
The Effect of the Study of Latin Upon Character Traits <i>Elizabeth B. Meek</i>	241
Teaching International-Mindedness in the Social Studies <i>Don W. Campbell and G. F. Stover</i>	244
The Teaching of Courtesy in the Junior High School <i>Alice K. Milson</i>	249
Workbook Method or Oral Discussion? <i>Elmer W. Cressman</i>	250
Plane Geometry and Character Education <i>Grace E. Allen</i>	254
The Results of the Incidental Method of Instruction in Char- acter Education <i>F. R. Kniss, E. K. Robb, and E. A. Glatfelter</i>	259
✓ The Effect of Athletics on Certain Character Studies <i>J. L. Hackenberg, E. B. Yeich, and L. A. Weisenfluh</i>	264
Summary of the Penn State Experiments on the Influence of Instruction in Character Education <i>Charles C. Peters</i>	269
Book Reviews . . . . .	273

# MOTION PICTURES AND YOUTH

## The Payne Fund Studies

W. W. CHARTERS, *Chairman*

The series, in nine volumes, will set forth the results of the investigations made by the group of university psychologists, sociologists, and educationists under the auspices of the Motion Picture Research Council and with the assistance of the Payne Fund.

The scientific studies reported in these volumes were designed to form a series to answer the following questions: What sorts of scenes do the children of America see when they attend the theaters? How do the *mores* depicted in these scenes compare with those of the community? How often do children attend? How much of what they see do they remember? What effect does what they witness have upon their ideals and attitudes? Upon their sleep and health? Upon their emotions? Do motion pictures directly or indirectly affect the conduct of children? Are they related to delinquency and crime, and, finally, how can we teach children to discriminate between movies that are artistically and morally good and bad?

### COMBINATION VOLUMES:

- Motion Pictures and Youth: A Summary*, by W. W. Charters, Director, Bureau of Educational Research, Ohio State University. *Combined with*  
*Getting Ideas from the Movies*, by P. W. Holaday, Indianapolis Public Schools, and George D. Stoddard, Director, Iowa Child Welfare Research Station. \$1.50  
*Motion Pictures and the Social Attitudes of Children*, by Ruth C. Peterson and L. L. Thurstone, Department of Psychology, University of Chicago. *Combined with*  
*The Social Conduct and Attitudes of Movie Fans*, by Frank K. Shuttleworth and Mark A. May, Yale University.  
*The Emotional Responses of Children to the Motion Picture Situation*, by W. S. Dysinger and Christian A. Ruckmick, Department of Psychology, State University of Iowa. *Combined with*  
*Motion Pictures and Standards of Morality*, by Charles C. Peters, Professor of Education, Pennsylvania State College. \$2.00  
*The Content of Motion Pictures*, by Edgar Dale, Research Associate, Bureau of Educational Research, Ohio State University. *Combined with*  
*Children's Attendance at Motion Pictures*, by Edgar Dale.

### SINGLE VOLUMES:

- Children's Sleep*, by Samuel Renshaw, Vernon L. Miller, and Dorothy Marquis, Department of Psychology, Ohio State University. \$2.00  
*Movies and Conduct*, by Herbert Blumer, Department of Sociology, University of Chicago.  
*Movies, Delinquency, and Crime*, by Herbert Blumer and Philip M. Hauser, Department of Sociology, University of Chicago.  
*Boys, Movies, and City Streets*, by Paul G. Cressey and Frederic M. Thrasher, New York University.  
*How to Appreciate Motion Pictures*, by Edgar Dale, Research Associate, Bureau of Educational Research, Ohio State University. \$1.20

THE MACMILLAN COMPANY

60 FIFTH AVENUE

NEW YORK, N. Y.







# The JOURNAL of EDUCATIONAL SOCIOLOGY

*A Magazine of Theory and Practice*

---

VOL. VII

DECEMBER, 1933

No. 4

---

## EDITORIAL

During the summer of 1932, about half of the members of a seminar on experimentation in education conducted at Pennsylvania State College agreed to undertake coöperative controlled experiments on character education during the ensuing winter. This issue of THE JOURNAL is devoted entirely to the presentation of the findings of this set of experiments. The investigations deal almost exclusively with a single phase of the subject—the influence of *instruction*, of one kind or another, upon character development. Sixteen different investigators participated in the experiments, employing thirty pairs of matched groups and making one hundred and eighty measured comparisons between mean attainments of groups given some type of instruction headed towards character-development objectives and otherwise equal groups not so instructed. This is by far the largest mass of scientific evidence now available on the question of the potency of moral instruction in modifying conduct. In fact, the previous controlled experiments dealing with this topic have been so few and so small in scope that we may say the question has hitherto been nearly untouched. Even the thirty experiments constituting this set are enough only to scratch the surface, so many-sided and difficult of quantitative attack is the question. But it is hoped that the results herein presented will open the field in a stimulating manner and that they will provide some preliminary indications of the trends to be expected from more exhaustive research.

C. C. P.

## THE POTENCY OF INSTRUCTION IN CHARACTER EDUCATION

CHARLES C. PETERS

Perhaps the elders of every age are more or less troubled about the "morals" of the oncoming generation. The untamed youth seems to them wayward, irresponsible, and contemptuous of the "tried and proved" customs to an extent that causes alarm for the safety of the future. But at certain periods of especially hurried transition this uneasiness becomes more than customarily acute and there is an unusually diligent search for ways of "training the rising generation in character." The present seems to be one of these times. Character education has, therefore, swung to the center of the stage in educational discussion. All sorts of proposals are being made for reaching this objective and many different means are being put into practice.

*But can character be improved by teaching?* Will our plausible-looking programs actually produce desirable outcomes? Or shall we be obliged, as we look back upon our efforts from the future, to admit that, although our ambition was pathetically earnest, our means were foolishly conceived? We in America have in general vast faith in "education." Whenever we find some weakness in our social order we bethink ourselves of "education" through the schools as the way to remedy it. But it is probable that we greatly overestimate the potency of formal education as a means of affecting conduct. It is probable that our civic education, our cultural education, and even our vocational education make far less difference in the functioning abilities of the persons to whom they have been applied than we are in the habit of believing. Controlled experiments on the functioning of school instruction have been disconcertingly disillusioning to educational optimists. It may be well that this same thing will prove to hold for instruction intended to improve "character." It behooves

us, therefore, at the beginning of our efforts to take measured stock of what we can accomplish by instruction. The series of experiments described in this issue of THE JOURNAL undertakes such evaluation with regard to certain representative programs of instruction.

In this series of articles we are using the term "character" very loosely. We should certainly not wish either what we include or what we omit to be taken as any contribution towards defining the scope of "character education." Character consists of an aggregate of habits, attitudes, and functioning philosophies of life of which only illustrative ones come within our set of experiments. Within "character," but by no means exhausting its connotation, is what we more narrowly call "morality"; that is, conformity with the *mores* of the societies to which the interpreter belongs.

As far as character is an acquired thing (which it is chiefly, if not entirely), it has two intertwined sources—imitation of others and trial-and-error experience on the part of the subject himself. Of these the former is by far the most frequent source. The *mores* are transmitted almost entirely in this manner and certainly "social suggestion" and "social radiation" are very powerful in molding all attitudes, tastes, and appreciations; and they make vast contributions in the shaping of ideals and philosophies of life. But the socially transmitted ways are tested in the individual's own experience and somewhat reshaped to fit reality as he finds it; and, especially in the reflection of the more philosophically tempered members of society, they may be radically and profoundly reshaped.

In consequence of these sources of character it is clear that the major factor in education for character formation must be social pressure from the groups to which the individual belongs, mostly unconscious pressure to which the individual yields little by little without knowing it. Nevertheless, it is within the power of educational executives (including teachers) somewhat to shape and direct, or at least to select, the pressures.

But it is in the learner's own experience that these socially proffered techniques are tested, refined, and personally assimilated. Experience, too, can be directed in school. It is, indeed, the sole function of a school so to set the stage that pupils will get fruitful experiences as rapidly and economically as possible. So character may be shaped through the participations of pupils in class activities, through the clubs and projects of the school, through organizations for pupil government, through the give-and-take of conversation and other forms of social life, and through every sort of dynamic experience in play and work. In all of these activities pupils try various techniques and select those that they find successful, or they observe the results of others' activities and accept for themselves those ways that they observe to be effective while rejecting those that appear "wrong." These accepted ways they build into their habit systems, their ideals and attitudes, their convictions and philosophies of life.

But the experience by which the socially proffered techniques are tested and assimilated need not necessarily be of the overt motor type. Thinking, too, is a kind of acting. When a person deliberates, he is trying out alternative ways of responding to a situation just as he is doing in trial-and-error experience, except that he is confining his trials to incipient acts carried in mental imagery and perhaps tagged through the aid of language. So reflection may be a substitute for direct trial-and-error experience after one has had enough overt experience to afford him types of known sequences upon which to draw. Just as one may watch others acting and learn from their successes and failures, so he may follow in imagination the conduct of characters narrated in anecdote, in literature, or in history. Thus, there is, besides social pressure operating through social suggestion and social radiation, and besides direct experience operating through personal trial and error, a third means of acquiring those readinesses to respond in which character consists—by vicarious experience in reflection, discussion, and listening critically to narrations of the experiences of others.

It is the management of this third item that we name "instruction" when we use that term in its strict sense. To instruct is to manage the situation in such a way that a pupil shall have ideas come before his mind for consideration that promise effective ways of achieving the ends he wishes to attain. Sometimes this takes the form of setting a concrete model and directing attention to its salient features, which we call demonstrating. Sometimes it has the form of directly proposing, which we call lecturing. Sometimes it involves marshalling a mass of ideas by developmental questioning. Sometimes it encourages deliberation by suggesting alternative possibilities or inducing the pupil to assemble alternative possibilities by his own systematized search. And sometimes it favors the presence of many alternative proposals out of which choice may be made by setting the stage for group discussion. But in all legitimate instruction it is the pupil himself who must accept, out of the proffered possibilities, those that he feels will work. Thus instruction is a very different thing from authoritatively telling a pupil "what's what" and expecting him passively to receive this.

Therefore, instruction is really not fundamentally different from learning through experience or from imitation. In learning from personal experience one accepts those ways of responding that prove fittest by his own direct trials. As he accumulates experience he is able to substitute imagined experiences for real ones and hence to make choices on the basis of deliberation. As he watches others he puts himself in their places and profits from their experiences vicariously, *i.e.*, imitates them. As his stock of experience becomes enriched and he gets effective command of language, he can live through these experiences of others when narrated, or even when put into the form of abstract generalizations, and can thus with far greater rapidity avail himself vicariously of the findings of the experiences of others. If, while living in this realm of the abridged actions that we call ideas, he can have the aid of a guide whom we call a teacher, to help him find revelant leads



and to see the abortive consequences of false leads, he can make the ideational substitute for overt experience so much the more effective. If he can match wits with his peers in group discussion while testing the probable fitness of proposed lines of action, this vicarious living is likely to be still more effective. Thus, between learning by direct experience and learning by instruction there is no sharp break; the latter is only more schematic and symbolic than the former.

It is, consequently, a plausible hypothesis that school instruction may be made a potent means of character formation. Is this hypothesis true? If so, we as educators are in a happy condition, for instruction is cheap and easily managed as compared with the total of the direct experiences of children. If it is not, we are in an awkward position, for instruction constitutes the major portion of the strategy of all conventional schools, if not of all schools. To test the truth of this hypothesis was the purpose of the set of experiments described in this issue.

Besides this question of the functioning of *instruction* in the shaping of character, there are several other questions relating to the possibility of purposive training for character, answers to which should be sought through scientific research:

1. What are the indirect contributions to character education from different methods of teaching school subjects? Miss Allen's experiment, in this series, is suggestive of possibilities here.

2. To what extent can different school subjects be made to contribute to character education by reason of certain emphases within them? To this possibility the studies of Miss Meek and Messrs. Campbell and Stover are pertinent.

3. Do extracurricular activities contribute, or can they be made to contribute, to the development of desirable character traits? This is a question on which there is much argument but extremely little experimental evidence. The only material this series has on it is the very inconclusive set of experiments on athletics by Hackenburg, Yeich, and Weisenfluh.

4. Do the disciplinary policies and practices of the school significantly affect personality traits? We have no evidence on this. We are hoping for an opportunity to attack this problem in the following way: From a large school system select several hundred junior-high-school pupils who have come up through the grades under teachers who are more



of martinets than the average and several hundred others who have come up under teachers who give pupils more than the average amount of room for freedom and initiative; then compute biserial or tetrachoric coefficients of correlation between strictness in discipline and each of a number of measurable character traits.

5. How do such nonschool educational agencies as the movies affect character development, and in what forms can these agencies be made to contribute useful values? On the former part of this question some evidence is supplied by the series of Payne Fund studies now being published by The Macmillan Company.

All of the experiments involved in our series are of the matched-group form. In each case a number of subjects were given a certain type of instruction and an equal number were used as a control group. The members of these two groups were matched, pair by pair, on one or more criteria for probable ability to improve in the experimental trait. This matching of groups by individual pairs not only makes the mean ability score the same for both groups but also makes the shape of the distributions the same at all points. Any matching criterion is valid that gives promise of high correlation with ability to make progress in the trait towards which the experimental factor is directed. Ordinarily, matching simultaneously on a number of criteria, each of which is correlated with ability to learn in respect to the trait in question, but which are not highly correlated with one another, gives better matched groups than pairing on a single criterion, but it also renders difficult the making of pairs. Probably the best scheme of pairing is one that involves some measure of rapidity of learning—one of the quotients—plus measurement of initial status in the experimental trait, for at least three reasons: (1) attainment to date is likely to be highly predictive of learning ability in the trait considered; (2) matching on the basis of initial attainment places the two mates at about the same position on the learning curve, and position on the learning curve at the beginning of the race has much to do with the prospect of improvement; and (3) matching on initial scores with which final scores are to be compared is likely to place together mates who have experienced similarly signed errors of measurement when the pairing is

also based on the second criterion suggested above. In addition to being matched for learning ability, both groups in each of our experiments were, of course, treated exactly alike except in relation to the experimental factor.

When a scientist has found an apparent law he always wishes to know with what degree of assurance he may depend upon it. Consequently, we wish to know the reliability of our findings in educational experimentation. The conventional formula for the reliability of a difference between two means is:

$$\sigma_{m_1-m_2} = \sqrt{\sigma_{m_1}^2 + \sigma_{m_2}^2 - 2r_{12}\sigma_{m_1}\sigma_{m_2}}$$

Recognizing that  $\sigma_m$  equals  $\frac{\sigma}{\sqrt{n}}$ , and that the  $n$  is the same for both groups when the members are arranged in pairs, we have:

$$\sigma_{m_1-m_2} = \sqrt{\left(\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n} - 2r_{12}\frac{\sigma_1\sigma_2}{n}\right)}$$

But in much experimental work with matched groups the third term, containing the  $r$ , is illegitimately omitted, resulting in a standard error that is too high. This is sometimes done because of ignorance of the true formula but often on account of the labor involved in computing the coefficient of correlation. Fortunately there is a very much simpler formula that gives identically the same results as the three-term one above which, for some strange reason, workers in statistics have almost completely overlooked. There are several ways of developing this simple formula, but we shall get it by making the conventional formula, given above, our starting point.

One of the forms of the Pearson product-moment correlation formula is:

$$r_{12} = \frac{\sigma_1^2 + \sigma_2^2 - \sigma_d^2}{2\sigma_1\sigma_2}$$

where the  $d$  is a difference between paired scores in the two arrays. Let us substitute this value for  $r$  in the second

of our reliability formulae just given. Doing the indicated cancelling we shall have:

$$\sigma_{m_1-m_2} = \frac{\sqrt{(\sigma_1^2 + \sigma_2^2 - 2(\frac{\sigma_1^2 + \sigma_2^2 - \sigma_d^2}{2\sigma_1\sigma_2})\sigma_1\sigma_2)}}{n} = \sqrt{\frac{\sigma_d^2}{n}} = \frac{\sigma_d}{\sqrt{n}}$$

Thus, in order to obtain the standard error of the difference between the means we take the differences between the end scores of paired individuals, find the standard deviation of this set of paired differences, and divide that by the square root of the number of pairs. Although this requires the computation of no coefficient of correlation, it takes full cognizance of the force of any element of correlation that is present.

In most of our experiments the results are in the form of the differences between *gains* by the two groups between initial measurements and final ones. The conventional formula for the standard error of the difference between mean gains is a very long and complicated one, consisting when correctly written of ten terms as compared with three in the one for end differences, and six of these terms involve the six possible intercorrelations among the four arrays. But I have shown elsewhere (in a book on statistics soon to be published) that we have an exact equivalent of this cumbersome formula in a very simple one parallel to the one just given for differences between end scores:

$$\sigma_{m_{g_1} - m_{g_2}} = \frac{\sigma_{dg}}{\sqrt{n}}$$

That is, we subtract an individual's initial score from his final score to find his gain; we similarly find the gain made by his mate; we take the difference between these two gains (which we call *dg*), find the standard deviation of the array of these differences in gains, and divide this standard deviation by the square root of the number of pairs. The procedure for end scores is illustrated in Table IV on page 242 and that for gains in Table II on page 235.

All the standard errors in connection with our experi-

ments are computed by these methods. For the benefit of relatively lay readers these are usually labeled S.E.diff. in our tables.

How great must a difference be in order to be significant? It is often said that it must be three times its standard error. But to say that is to appeal to a kind of magic. As a matter of fact there is no precise ratio at which a difference becomes significant. It is all a matter of odds against reversal of the advantage. When a difference is three times its standard error the chances are, assuming a normal distribution of differences from successive samples, 740 to 1 that the true difference is in the indicated direction; if the ratio is 2, the chances are 43 to 1, and if the ratio is .8, the chances of a true difference in the same direction are 3.7 to 1. When a ratio of three is demanded the great majority of differences turn out to be "not statistically significant" and the implication is left that the two procedures are of equal value even though the chances may be several hundred to one that continued experimentation would show an indicated one superior to the other. Personally I should like to bet on the stock market with the chances even three or four to one in my favor, and similarly I am willing to bet on a method of improving character while we await further experimental evidence with the odds not much greater.

Another important consideration is *the direction of the differences in duplicated experiments*. If several experiments give differences in the same direction the reliability is greatly increased. It is a well-known principle in the mathematics of probability that if the probability of the occurrence of a given event is  $p$  when one condition obtains and  $q$  when another condition obtains, that probability is  $p$  times  $q$  when both conditions obtain. By this law if the probability of having obtained a difference of a certain size in favor of an experimental factor when the true difference is on the other side is  $1/4$  in one experiment and  $1/6$  in a second experiment, it is the product of these two, or only  $1/24$ , that a difference would have been obtained of

these sizes on this same side in both the trials if the true difference did not lie on that side. This same principle would hold for any combination of experiments, although a different principle must be applied when differences lie on opposite sides of zero.

This is, of course, true only if the experiments are independent of one another. If they involve the same pupils or the same teachers, but different measures of success, so that there is some element of correlation present, we cannot simply multiply together the probabilities. Nevertheless, under all circumstances except perfect correlation (perhaps never present) a set of differences pointing prevailingly in the same direction indicates much higher reliabilities than those of the separate trials. In many themes on which we experiment, differences are real but inherently small. The summary of the experiments of this set suggests that, for systematic instruction, the true differences average about four tenths of a standard deviation. I have determined that, neglecting the correlation factor between gains (likely to be very small), and ignoring the slight difference between the standard deviation of a single array and that of the two matched arrays combined, it would require 113 pairs of subjects in an experiment showing this difference to reach a ratio of three or more in half the trials and 153 pairs to reach such ratio in two thirds of the trials. Such groups are not attainable as single groups under ordinary school conditions. X

In our next article we shall justify the use of ratings as measuring devices, upon which we have leaned heavily in this set of experiments. In the following articles we shall set forth the experimental findings in as much detail as space permits. After these presentations of details, I shall summarize the findings from the set of experiments as a whole, putting these in a form that is readily comparable for all of the nearly two hundred experimental comparisons, and draw the indicated conclusions.



## THE RELIABILITY AND VALIDITY OF ESTIMATES (RATINGS) AS MEASURING TOOLS

JAMES C. SWAB AND CHARLES C. PETERS

One of the most serious obstacles to controlled experimentation in character education, as well as in certain other areas, is the lack of suitable measuring instruments. Verbal tests have been highly developed during the past quarter of a century, and certain types of non-verbal performance tests have also been brought to a high state of perfection within certain areas. But verbal tests have distinct limitations; they reveal chiefly informations, judging abilities, and perhaps preferences and attitudes. But practical conduct in life situations may not agree, at least completely, with these declared informations, judgments, and preferences. Performance tests, as we know them, must usually be forced in such an artificial manner in order to yield objective scores as to make it impractical to use them outside of a specially prepared laboratory. Within the past few years educational and psychological research workers have been trying out ratings based on free estimates as measuring tools and have been agreeably surprised at the reliabilities and validities shown when averages from a number of judges were involved. This study has as its object an investigation of the reliability and validity of estimates. To show the evidence it contains is particularly necessary at this point because most of the articles in this magazine make use of pupil-and-teacher estimates as their chief measuring devices.

This study involved 30 pupils in the seventh grade of a small Pennsylvania school system and 34 pupils in the eighth grade. Since these were all the pupils in those grades, they each had the opportunity to know one another very intimately. For some of the traits dealt with in the investigation the objective facts were known, so that we had validity criteria for the ratings relating to them, while for others we had no such criteria. We could, therefore, in-



investigate the reliabilities of the estimates for all of the traits studied but the validities of the estimates for only those for which we had objective facts.

The traits involved in the study were the following: honesty, courtesy, brightness, ability in arithmetic, height, and age. Honesty was used in the sense of the following definition: Honesty is that quality of man that shows him fair and truthful in speech; above cheating, stealing, misrepresentation, or any other fraudulent action. Courtesy is showing consideration for others; politeness; favor as distinguished from right. Brightness was evidenced to the pupils who constituted the judges by ability to answer or to recite well in class in all school subjects. For purposes of a validity criterion it was determined by scores on an intelligence test. The other terms—arithmetic grades, age, and height—were used in the conventional sense.

Each pupil in a section was given a pack of cards containing the names of all the members of the section (grade). The pupils were asked to group these names, representing the pupils of the class, into five stacks: tallest, tall, average, short, shortest; or oldest, old, average, young, youngest; or whatever else was the trait being ranked. They were then asked to complete the rankings within each pile so that all the pupils in the room would be ranked from the highest to the lowest. In making these rankings the pupils had in mind the definitions given above. From records of the actual facts regarding the pupils on those traits for which we had validity criteria, the cards were also ranked and the ranks recorded.

For the purpose of determining the validity of the estimates a composite rank was obtained for each pupil by averaging the ranks assigned him by his mates, reranking these composite scores according to relative size, and then computing the coefficient of correlation between ranks in estimates and paired ranks according to the actual measurements of the trait in question. The correlations were computed by the Spearman ranks method and the rho's translated into corresponding  $r$ 's by means of tables.

Reliability coefficients were computed in two different ways, in order that we might check the results of the two methods against each other.

1. A score was obtained for each pupil from a random half of the raters and another from the other half, and the scores from these two halves correlated by the Pearson product-moment method. Since this gave the reliability of the average of the estimates of only half of the judges against another half while we needed that of the whole set against another set of equal size and character, we inferred the latter by application of the Spearman-Brown formula:

$$r_{\bar{a}\bar{a}} = \frac{2r_{\frac{a}{2}\frac{a}{2}}}{1 + r_{\frac{a}{2}\frac{a}{2}}}$$

That is, we divided twice the  $r$  between the scores from the halves by 1 plus this  $r$ .

2. We obtained the average intercorrelation among the ranks for the 30 judges in the seventh grade, or the 34 judges in the eighth grade, by the following formula:<sup>1</sup>

$$r_{11} = 1 - \frac{a(4N+2)}{(a-1)(N-1)} + \frac{12\S S^2}{aN(a-1)(N^2-1)}$$

In this the  $a$  is the number of judges, the  $N$  is the number of pupils ranked (which in this particular case was the same as the number of judges), the  $S$  is the sum of the ranks for a particular pupil assigned by all the judges, and the  $\Sigma S^2$  the aggregate of the squares of these pupil sums for all the pupils in the class. These average intercorrelations ranged, for the various traits in the two grades, from .412 to .839 and showed the extent of agreement, on the average, between the rankings of any two judges.

Our concern was not, however, with the extent of agreement of one judge with another but rather with the extent of agreement to be expected between *the average rankings by the whole set of judges* and the average by another set of the same size that might in the future be drawn from

<sup>1</sup>The proofs for all of the formulae in this article are given in T. L. Kelley, *Statistical Method* (New York: The Macmillan Company, 1923), pp. 205-218, and in C. C. Peters, *Motion Pictures and Standards of Morality* (New York: The Macmillan Company, 1933), pp. 278-283.

the same sort of population. This  $r$  between the average from our whole set of judges ( $a$  in number) and the average from a similar set can be predicted by the Spearman prophecy formula:

$$r_{a\bar{a}} = \frac{ar_{11}}{1 + (a-1)r_{11}}$$

where the  $r_{11}$  is the average intercorrelation found by the preceding formula and  $a$  is again the number of judges.

We may also infer the extent to which the average from our  $a$  judges would agree with the average from an indefinitely large number (the so-called "true" estimate) as follows:

$$r_{a\infty} = \frac{ar_{11}}{\sqrt{ar_{11} + (a^2 - a)r_{11}^2}}$$

The results for these several computations for the various traits for the two grades are displayed in the accompanying table. The first row across (1) gives the average intercorrelation among the judges; the second line (2) shows the inferred reliability correlation for the whole set of judges against a second similar set by way of the average intercorrelations; the third line (3) shows the correlations between the average ratings and the "true" ratings; the fourth line (4) gives the  $r$  between the ratings averaged from random halves of the judges; and the fifth line (5) gives the total reliability coefficient by way of the Spearman-Brown formula—a value that should be parallel to line 2. The other lines are self-explanatory.

Inspection of lines numbered 2 and 5 in the table shows extremely high reliability coefficients. In no case does the coefficient fall below .946 and prevailingly the  $r$ 's are around .98. It is only rarely that objective verbal tests reach as high as this. How closely the average estimates from the 30 or the 34 judges agree with the "true" estimates is revealed in lines numbered 3. These  $r$ 's fall not far short of unity. A second feature worth noticing is the very close agreement between results from the two methods

TABLE I  
SUMMARY OF RELIABILITY AND VALIDITY COEFFICIENTS  
Seventh Grade, Reliability

	Honesty	Courage	Age	Brightness	Arithmetic	Height
1. $r_{II} = 1 - \frac{a(4N+2)}{(a-1)(N-1)} + \frac{12\S S^2}{aN(a-1)(N^2-1)} =$	.412	.447	.557	.604	.673	.839
2. $r_{aa} = \frac{ar_{II}}{1 + (a-1)r_{II}} =$	.960	.960	.974	.978	.985	.993
3. $r_{aa} = \frac{ar_{II}}{\sqrt{ar_{II} + (a^2 - a)r_{II}^2}} =$	.979	.955	.985	.989	.992	.996
4. $r_{\frac{a}{2}} = \frac{a}{2} =$	.940	.898	.970	.953	.952	.989
5. $r_{aa} = \frac{2r_{aa}}{1 + r_{\frac{a}{2}}} =$	.970	.946	.985	.975	.978	.993
6. $P = 1 - \frac{6\S D^2}{N(N^2-1)} =$	.....	.....	.508	.327	.756	.962
7. Corresponding $r =$	.....	.....	.525	.340	.770	.965
8. $P.E. P = \frac{.7063(1-P^2)}{\sqrt{N}} =$	.....	.....	.09	.16	.05	.01

Seventh Grade, Validity

Eighth Grade, Reliability

	Honesty	Courage	Age	Brightness	Arithmetic	Height
1. $r_{II} = 1 - \frac{a(4N+2)}{(a-1)(N-1)} + \frac{12\S S^2}{(aN-1)(N^2-1)} =$	.704	.506	.415	.640	.575	.771
2. $r_{aa} = \frac{ar_{II}}{1 + (a-1)r_{II}} =$	.987	.972	.960	.984	.979	.991
3. $r_{ao} = \frac{ar_{II}}{\sqrt{ar_{II} + (a^2 - a)r_{II}}} =$	.993	.985	.980	.992	.990	.997
4. $r_a \frac{a}{2} =$	.962	.952	.925	.971	.963	.980
5. $r_{aa} = \frac{2r_{aa}}{1 + r_{II} \frac{1}{2}} =$	.986	.975	.958	.986	.986	.990

Eighth Grade, Validity

	....	....	.594	.199	.834	.825
6. $P = 1 - \frac{6\S D^2}{N(N^2-1)} =$	....	....	.594	.199	.834	.825
7. Corresponding $r =$	....	....	.610	.210	.845	.840
8. $P.E.P = \frac{.7063(1-P^2)}{\sqrt{N}} =$	....	....	.08	.11	.04	.04

of determining reliability, as indicated by lines numbered 2 and 5.

The validities, shown at the lower part of each section of the table, are of especial concern to us, for they show the extent to which the estimates conform to the objectively determined truth. For height the validity correlation is .965 in the seventh grade and .840 in the eighth grade. These good validities are doubtless to be attributed in part to the fact that height is a readily observable trait, and the greater validity in the seventh grade than in the eighth may be due to greater heterogeneity in the former grade than in the latter.

For age the correlation expressing the degree of agreement between average estimate and objective fact is .508 in the seventh grade and .594 in the eighth. While these  $r$ 's are not very high, they must be viewed in the light of the homogeneity of the groups, for it is well known that small  $r$ 's where the variability in either or both of the groups compared is slight are equivalent to much larger ones where the variabilities are greater. The semi-interquartile range of ages was only nine months in the seventh grade and only three months in the eighth.

The arithmetic grades the pupils should receive were estimated by the judges in a way that correlated with the grades later given by the teacher: .770 in the seventh grade and .845 in the eighth. Both these correlations may be considered high when we remember the possibility of a certain lack of reliability in the grades themselves and also the lack of perfect validity in the grades. If these  $r$ 's could be corrected for attenuation, they would probably not fall far short of unity.

There remain the validity correlations for estimates of brightness. These were .340 in the seventh grade and .210 in the eighth grade. At first these coefficients look unreasonably low, but their lowness may be explained by possible lack of validity in the intelligence-test scores themselves. It must be remembered that intelligence-test scores correlate with teachers' grades or with other objective meas-



ures of scholarship only from about .20 to .60, averaging perhaps .38 or .40. Even these objective measurements of achieved scholarship may be more narrow in scope than the thing the pupils meant by brightness. Furthermore, the range of talent was not very great at this level for a single grade and was probably less in the eighth grade than in the seventh. If our validity correlations could be corrected for these faults in the criterion, they might well be satisfactorily high.

A number of other investigations made at Penn State, in which the reliabilities of ratings were involved as a by-product, agree with this one in showing high reliabilities and high validities for averages from ratings. Campbell obtained a reliability coefficient of .985 by having 39 members of a social fraternity rate one another on "personal culture," Merrill got, by the split-halves method for the experiment reported in this series, a reliability coefficient of .935 in the fall ratings and .823 in the spring ratings, while Eichler's reliability coefficient for his ratings on leadership by pupils was .964. In the evaluation of motion pictures by committees of five members we obtained twenty-six reliability coefficients ranging from .76 to .98, usually in the .90's. Twenty reliability coefficients were computed on evaluations of the moral quality of certain described bits of conduct in connection with our study of motion pictures and standards of morality by groups of from 18 to 50 members each. These ranged from .796 to .981 and averaged .933. When groups of 187 members were made up by consolidating the smaller groups, the *r*'s for the four types of themes were .987, .994, .990, and .983. From estimates of the pleasure-giving values of items in chemistry education Wray got sixteen reliability coefficients ranging from .751 to .951 when groups of from 9 to 36 members each were used, .943 when a group of 142 members was used, .953 for a group of 176 members, and thirteen other such *r*'s ranging from .910 to .956 from other groups of this same order of size. In a similar study dealing with psychology Lick obtained a reliability coeffi-

cient of .94 from 100 judges. From repetition of ratings after a long interval Wray obtained  $r$ 's of .957 and .980 from a group of 38 college seniors.

Suggestive of both the validity and the reliability of the ratings, we have obtained regularly rather high agreements among different groups rating the same objects. Eichler's pupil ratings on leadership correlated with teacher ratings .77. Glatfelter's pupil ratings showed the following  $r$ 's with those by teachers rating the same persons for the same traits: coöperation .795; courtesy .754; industry .829; loyalty .779; dependability .779. Forty intercorrelations among different groups on the evaluation of the moral quality of described acts in our motion-picture study averaged .838. Wray calculated 43 intercorrelations among diverse groups as to the values found in certain items of chemistry education and found them to average .736 (uncorrected for attenuation, as all of them are which are quoted here). Himes found correlations between boys and girls in the ratings on pleasure values in biology to range between .73 and .87, and to average .81. In view of the fact that real differences among the groups would bring these  $r$ 's somewhat below unity even if the measures were perfect, such high coefficients of correlation could not be obtained unless the ratings as handled had both good validity and good reliability.

Investigators other than those at Penn State have experienced similarly satisfactory results from averages of ratings.

In view of all the evidence accumulated during the past few years no one can any longer deny to ratings a place beside objective verbal tests as dependable measuring devices—uniquely valid for measuring certain types of functioning conduct in normal life situations.

## CAN SOCIAL LEADERSHIP BE IMPROVED BY INSTRUCTION IN ITS TECHNIQUE?

GEORGE A. EICHLER AND ROBERT R. MERRILL

One of the most boasted objectives of practically all types of schools is training for leadership. This objective is, however, customarily not clearly defined and usually fuses more or less vaguely two elements: (1) outstanding technical expertness of a type that will get the individual who possesses it looked to and sought as an authority in his field; and (2) the attributes and techniques that enable one to set standards of conduct for others and particularly that enable him to command a following among others—to direct and control other individuals or groups. We shall call this latter type of leadership, the actual management of other individuals and groups, *social leadership*. When hard pressed, most educational policy makers, especially in the higher institutions, will admit that it is the former that they chiefly mean when talking of education for leadership. Yet the latter is also extremely important in society, especially in a democratically organized society. As yet schools seem to have consciously done little about it and their attainments in respect to it appear to be as meager as their efforts.

If ways could be found for improving among students in training the ability to lead others by effective techniques towards socially desirable ends, the educational contribution thereby made would be of inestimable importance. Can social leadership be improved by systematic training? Since it is conditioned by the employment of certain techniques, can at least a partial mastery of these techniques be developed in pupils by instruction? Can the basic skills involved in leadership be developed by guided practice? Or can a functioning leadership be, perhaps, improved by a combination of guided practice paralleled by a theoretical consideration of techniques? To secure an

answer to the first of these questions was the purpose of the two experiments described in this article. We hope, in a continuation of the experiment, to find an answer to the second and third of the questions.

#### PROCEDURE

The parallel group method of experimentation was used. Student ratings of one another on leadership were used in both studies. In study A,<sup>1</sup> a mimeographed list of 72 members of a sophomore class was handed to this class and the members were requested to rank every classmate on leadership. In study B, students not used in the experiment were given ballots in the form of 3x5 cards on which were names of some of their classmates. They were asked to rate their classmates on leadership on a five-point scale by encircling the proper number. The encircled numbers for each particular student were added and the result divided by the number of cards marked for him. This gave a leadership index for the student. On the basis of this information, control and experimental sections were paired. Study B was carried on with a control and an experimental section in grades nine and twelve in two different schools in order to secure a check on the experiment. In both studies the experimental groups were taught lessons in leadership. This instruction in study A consisted of six forty-five minute lectures on leadership qualities and techniques. Instruction in study B consisted of eleven thirty-minute conferences on various qualities and techniques of leadership. In the case of study A, the instruction was given over a period of six weeks while in study B the instruction was spread over about seven months. The reliability coefficient of the ratings used for the measurement of progress in study A was .935 and in study B .964.

The table on page 235 gives the results for the twelfth grade in study B. On the left the data are given for the pupils of the experimental section and on the right, on the

<sup>1</sup>These experiments were carried out by R. R. Merrill in Youngsville, Pennsylvania, in 1931, and by G. A. Eichler in Northampton, Pennsylvania, in 1933. The former study will be referred to in this article as study A, and the latter as study B.

same horizontal lines as their mates, the pupils of the control section. The columns headed IR give the initial ratings; those headed ER, the end ratings; those headed G, the gains during the period of experimentation; while the DG column gives the differences in gains by paired mates.

TABLE II  
COMPARATIVE GAINS IN LEADERSHIP RATINGS BY INSTRUCTED AND UNINSTRUCTED  
PUPILS—STUDY B

Experimental Group				Control Group			
Pairs	IR	ER	G	IR	ER	G	DG
1	403	370	— 33	402	383	—19	— 14
2	390	423	+ 33	395	397	+ 2	+ 31
3	372	380	+ 8	378	374	— 4	+ 12
4	365	374	+ 9	366	395	+29	+ 20
5	352	347	— 5	351	363	+12	— 17
6	344	297	— 47	337	300	—37	— 10
7	334	200	—134	332	294	—38	— 96
8	327	369	+ 42	328	368	+40	+ 2
9	318	391	+ 73	319	248	—71	+144
10	277	250	— 27	276	245	—31	+ 4
11	263	289	+ 26	263	242	—21	+ 57
12	261	292	+ 31	257	246	—11	+ 42
13	248	203	— 45	248	198	—50	+ 5
14	239	273	+ 34	240	204	—36	+ 70
15	229	190	— 39	228	232	+ 4	— 43
16	221	239	+ 18	225	232	+ 7	+ 11
17	217	206	— 11	223	187	—36	+ 25
18	214	200	— 14	207	193	—14	0
19	202	177	— 25	204	223	+19	— 44
20	200	189	— 11	203	226	+23	— 34
21	190	208	+ 18	187	206	+19	— 1
22	150	150	0	156	160	+ 4	— 4
Mean	278	273.5	—4.5	278.4	268.9	—9.5	+5.

It will be noticed that the experimental group lost an average of 4.5 and the control group 9.5 which nets a difference of 5. in favor of the experimental group. The fact that both groups lost is not significant; in general it only indicates a different degree of leniency in rating at the beginning of the experiment from that at the end; it is only the comparative rating on the two groups that is important. We find the standard deviation of the differential gains to be 46.97. We are now interested to know the standard error of the difference of the mean gains.

By use of the formula  $\sigma_{\text{diff}} = \frac{\sigma_{d's}}{\sqrt{n}}$  we find  $\sigma_{\text{diff}}$  to be 10.

The difference between the means of the gains is 5, which divided by 10 yields .5 as the ratio between the standard error and the difference. This indicates that the chances are 2.2 to 1 that there is a true difference in favor of the instructed group. The results in the ninth grade were



strikingly similar to those in the twelfth grade, the chances being 2.2 to 1 that the true difference is above zero in favor of the experimental group.

The results in study B are summarized as follows:

Mean gain in score of experimental group.....	.976
Mean gain in score of control group.....	— .71
Difference between gains in favor of experimental group.....	1.686
$\sigma$ of difference of the means.....	1.178
Ratio of the diff. to $\sigma$ of difference.....	1.431
Chance of true difference being greater than zero in favor of the experimental group .....	12 to 1
Per cent of pupils who gained in score in experimental group..	54.8
Per cent of pupils who gained in score in control group.....	43.9

#### INTERPRETATION OF RESULTS

From a statistical point of view as ordinarily interpreted, the results obtained are far from significant, but they are all in the same direction which greatly strengthens the reliability. In view of the fact that progress in so complex a trait as social leadership is a very different matter from progress in the acquisition of simple motor skills, the results obtained are all that any one could anticipate, for no one would expect to make leaders overnight. If instruction is effective in making a noticeable difference in the short period of these experiments, we may hope to achieve considerable success by continuing proper instruction through the whole secondary-school period.

#### CONCLUSIONS

The results of the studies summarized seem to justify the following conclusions:

1. It seems possible to measure reliably leadership qualities by means of student ratings.
2. It is probable that leadership qualities can be measurably improved by direct instruction.



## THE EFFECT OF DIRECT INSTRUCTION

E. K. ROBB AND J. F. FAUST

Two experiments are described dealing with the possibility of improving ethical discrimination and moral conduct by systematic instruction in the senior high school on ethical problems. The experiments were conducted at Bedford by Mr. Robb and at Chambersburg by Mr. Faust.

### I. THE BEDFORD EXPERIMENT

Two sections of seniors were matched according to I.Q. as measured by the Otis Group Test, and socio-economic status as measured by the Sims Scale. Fifty-two students were included in the experiment, 26 of whom were in the control group and 26 in the experimental group.

The experiment was conducted in connection with the class in problems of democracy. In the control group the regular course of study in problems of democracy was followed throughout the term. In the experimental group this course was supplemented for an eight-week period with direct instruction on ethical problems with the use of Peters's *Human Conduct*<sup>1</sup> as a basic text. The instructor, Mr. W. Edward Sheely, encouraged class discussion of all problems related to the field of character education.

As a measure of the results of the experiment both groups were examined in moral knowledge and ethical discrimination by the use of the Kohs Ethical Discrimination Tests at the beginning and again at the end of the experiment. Teacher ratings were made for each individual at the beginning and at the end of the experiment by the aid of the Character Education Inquiry Conduct Record Sheet. Pupil ratings on the persons involved in the experiment were secured before and after the experiment on industriousness, leadership, honesty, courtesy, and loyalty. In taking the ratings five small cards were supplied to each pupil upon each of which he was requested to write the name of one

<sup>1</sup>C. C. Peters, *Human Conduct* (New York The Macmillan Company, 1918), 427 pages.

intimate acquaintance in his class. Ratings were taken separately on each of the traits. Since the senior class was small enough to permit pupils to know one another rather intimately, and since they all intermingled freely regardless of the sectioning involved in the experiment, the members of both sections chose students for rating indiscriminately from either section. When taking ratings on a trait, the pupils were instructed to arrange the cards of the ones whose names they had written in the order of proficiency in that trait. By recording these rankings with credits according to their rank order we had ratings for each pupil by a number of students for each trait. By averaging the ranks thus assigned to a student by all those who had rated him, a composite score for each pupil was computed.

The results of the Kohs Ethical Discrimination Test showed the difference between the means to be 4.77 with a standard error of 3.17, indicating chances of 14 to 1 that the true difference is in favor of the experimental group. From the teacher ratings the difference between the means was 5.31 with a standard error of 1.93, involving chances of 332 to 1 that the true difference is on the side of the experimental (instructed) group. From the data received from the pupil ratings, the difference between the means of the gains was found to be .03 in favor of the instructed group, with a standard error of .173 and chances of 1.35 to 1 that there is a true difference above zero in favor of the experimental group.

We therefore find for the Kohs Test and the teacher ratings reasonably significant differences, both pointing to an advantage for the group that had received systematic instruction in ethics. While the difference in gain as measured by pupil ratings is much too small to be individually significant, it points to the same direction as the other two. As far as the evidence from this small experiment goes, it suggests that moral discrimination of high-school seniors, and such moral conduct as that covered by our ratings, can be improved by systematic instruction in ethics.

## II. THE CHAMBERSBURG EXPERIMENT

Two homeroom groups, in the ninth grade, containing about thirty pupils each were selected to determine the effect that homeroom programs discussing moral problems would have upon the ethical judgment of the pupils. The pupils were matched directly according to chronological age, school grade, high-school course, and intelligence quotient. They were also matched in a general way on social characteristics, school activities, and school achievement.

The experimental group had homeroom programs one hour per week for eighteen weeks based on various attributes of character, including respect for authority, courtesy, honesty, loyalty, leadership, fair play, sex relationships, service, respect for matters sacred and religious, tolerance, dependability, and coöperation. These programs included a variety of activities, such as debates, discussions, dramatization, reports, observations, notebook projects, vocabulary drills, story-telling, and talks by teacher and principal. The control group had regular homeroom programs where character education was not stressed more than in the incidental way usual in such programs.

Both groups were tested at the beginning of the experiment and again at the end with tests designated as: (1) Character Attributes Test—rather puzzling moral situations somewhat fully stated; (2) Character Reaction Tests, Parts I and II—briefly described moral situations; and (3) Character Attributes Self-Rating Scale. The validity of the tests was established on the judgments of sixty-three adults from different occupations. In addition to these tests the two groups scored themselves at the conclusion of the experiment on the O'Reilly Character Analysis Chart.

A summary of the findings is given in the table on page 240. The negative sign favors the control group.

From Table III, on page 240, it may be seen that the experimental group excelled on two of the three judgment tests, while the control group made the greater gains on the other one. On the homemade self-rating scale the control

TABLE III

COMPARISON OF MEAN GAINS OF THE EXPERIMENTAL AND THE CONTROL GROUPS—CHAMBERSBURG EXPERIMENT

<i>Test</i>	<i>Difference in Mean Gains</i>	<i>Standard Error of Difference</i>	<i>Chances True Difference Same Side</i>
Character Attributes, complex problems	2.4	1.37	24 to 1
Character Reaction, I, briefly stated problems .....	— .90	1.33	3 to 1
Character Reaction, II, brief problems	.8	1.65	2.2 to 1
Self-Rating .....	— 1.6	.89	27 to 1
O'Reilly Character Analysis Chart...	2.8	2.2	9 to 1

group excelled and on the O'Reilly scale the experimental group made the greater gain. The test of moral judgment on which the control group made the greater gain consisted of short, categorical statements such as, "One should not stretch the truth"; "A leader is one who goes ahead because he has to"; "Dishonest persons are very desirable associates"; "We ought to be loyal to our superiors rather than our subordinates." Test 1, on which the advantage was most largely on the side of the experimental group, involved much more challenging problems stated at considerable length. The third of the judgment tests on which there was a small advantage to the experimental group had prevailing statements between these other two in complexity. It may be possible that the choppy statements of test 2 were too trite to challenge these pupils who had for a semester debated moral issues, or it may be that they involved half truths to an extent that elicited unexpected responses from pupils who had been practised in challenging moral issues, so that their scores at the end of the experiment suffered rather than benefited from the instruction. Or it may be that the outcome of this experiment should be interpreted as a draw, indicating no advantage from the discussion of problems of conduct.

*Conclusion.* While not entirely conclusive, these experiments suggest the possibility of slightly accelerating ethical development as measured by verbal tests, and functioning conduct as measured by ratings, through direct and systematic discussion of problems of conduct by early adolescents.

## THE EFFECT OF THE STUDY OF LATIN UPON CHARACTER TRAITS

ELIZABETH B. MEEK

Is the study of Latin developing better attitudes towards social situations, war, international attitudes? One of the most intangible school problems is the determination of the degree of success that the school has attained in developing desirable attitudes. Unfortunately, educational technique is not very well developed to enable one to measure with certainty such learning products. The chief difficulty is that when we measure understanding we are not measuring the related concrete behavior, and we do not know the correlation between the pupil's knowledge of right and wrong and his actual attitudes and conduct.

It is true that conduct is affected by other factors, such as the emotional factor, as has already been said, but it is probable that the disharmony, which sometimes seems to exist between knowledge and conduct, is due not to a real contradiction between the two, but to the fact that apparent knowledge is not real understanding. It may only be an imitative repetition of the opinion of others . . . . If education which is directed towards the improvement of conduct, then, can be shown to produce a substantial improvement in the comprehension of social situations by children, we have good reason to expect that it will produce an improvement in their conduct.<sup>1</sup>

With this assumption an attempt will be made to show how the incidental teaching of character traits through Latin has functioned in the experimenter's school. In the experimental group there were twenty pupils of the tenth grade who were taking Latin. For the control group twenty pupils of the tenth grade were found who were pursuing the same subjects under the same teachers, with the exception of Latin, and who could be paired with the experimental pupils. These forty pupils studied the same subjects under the same teachers in the first eight grades. In the ninth grade they had pursued the same subjects under the same teachers except that the twenty pupils in the experimental group had taken Latin in addition to the other subjects. They were paired on sex, chronological age, intelligence quotient, and on composite grade at the end of the

<sup>1</sup>Yearbook of National Education Association, Department of Superintendence, 1930, p. 756.



eighth school year. To determine the intelligence quotient the Terman Group Test of Mental Ability was administered.

To test the effect of Latin upon developing right attitudes towards social situations, Hill's Test in Civic Attitudes was used. Following are the scores obtained:

TABLE IV  
SCORES FROM HILL'S CIVIC ATTITUDE TEST

<i>Pair Numbers</i>	<i>C Score</i>	<i>E Score</i>	<i>Difference</i>	<i>Pair Numbers</i>	<i>C Score</i>	<i>E Score</i>	<i>Difference</i>
1	19	19	0	11	17	18	1
2	17	19	2	12	17	20	3
3	18	20	2	13	18	18	0
4	19	20	1	14	17	18	1
5	16	19	3	15	19	19	0
6	16	15	-1	16	18	15	-3
7	18	19	1	17	18	17	-1
8	16	17	1	18	17	18	1
9	16	20	4	19	12	14	2
10	19	17	-2	20	16	17	1
				<i>Mean</i>	17.15	17.95	0.8

Thus the difference between the means for the two groups is .8. The standard error of this difference is .37, making the ratio of the difference to its standard error 2.15 and involving chances of 62 to 1 that the true difference is in favor of the Latin group.

The result obtained shows that incidental teaching for character through Latin has had a positive effect. Although the difference is not very great, the period through which this teaching was given extended over only seven months.

To test the effect of Latin upon the attitude towards war, L. L. Thurstone's Attitude Toward War Scale, Number 2, Form A, was used. The following explanation enables one to interpret the individual scores as well as the average score of the group:

0 — 2.9	extremely militaristic	6.0— 6.9	mildly pacifistic
3.0— 3.9	strongly militaristic	7.0— 7.9	strongly pacifistic
4.0— 4.9	mildly militaristic	8.0—11.0	extremely pacifistic
5.0— 5.9	neutral position		

On this measure the mean score of the experimental group was 6.77 and that of the control group 6.30, showing a

difference of .47 in favor of the experimental group. The standard error of the difference is .20, giving a ratio of 2.35 between the difference and its standard error. The chances are, therefore, 105 to 1 that the true difference is on the same side.

The result obtained shows that incidental teaching of character building through Latin has had a positive effect; it has made the pupils more strongly pacifistic. This result is especially gratifying, as many critics adverse to the teaching of Latin have claimed that the reading of such literature as Caesar's *Gallic War* would have a tendency towards making pupils militaristic.

To test the effect of Latin upon international attitudes, Neumann, Kulp, and Davidson's Test of International Attitudes was used. A high score on the test indicates a tendency towards conservatism; a low score indicates a tendency towards liberalism. The results were worked up in precisely the same manner as in the two preceding cases. The Latin group made a mean score of 3.59 and the non-Latin group 3.89, a difference of .30 with a standard error of .094, a ratio of the difference to the standard error of 3.15 and chances of 1,200 to 1 of a true difference in the same direction.

The result obtained shows that incidental teaching for character building through Latin has had a negative effect towards producing a high score, towards making pupils conservative. It has, therefore, made them more inclined towards liberalism which is the desired effect. Of the three tests given this one shows the greatest difference between the two groups. This can very easily be understood when one thinks of the many opportunities presented by Latin history and literature for bringing to the pupils' attention the many fine character qualities manifested by the peoples of races and nations very different from our own.

Thus all three of our measures consistently suggest the possibility of developing desirable character traits by stressing them in connection with the teaching of Latin.

## TEACHING INTERNATIONAL-MINDEDNESS IN THE SOCIAL STUDIES

DON W. CAMPBELL AND G. F. STOVER

### I. THE CONNELLSVILLE EXPERIMENT (CAMPBELL)

The purpose of this study was to determine the possibilities of influencing high-school pupils to become more internationally minded by incidental teaching in economic geography. The investigation covered a period of eighteen weeks and was conducted by the writer in the high school of Conneltsville, Pennsylvania, an industrial community of 14,000 inhabitants. Due to the type of community in which the school is located, pupils of various nationalities were present. The four classes used were comprised of a heterogeneous grouping of sophomores, juniors, and seniors. Pupils were matched on I.Q. and on scores on the Neumann-Kulp-Davidson Test of International Attitudes. From 150 pupils originally tested, 80 were satisfactorily matched and were divided into control and experimental groups.

The teaching method employed in the instruction of both groups was the Unit Mastery Technique. Subject matter was unitized into economic regions of the world and was studied with the aid of mimeographed sheets which stated the objective, the reference readings, and the subproblems related to the major objective of the unit. The pupils were then given time during following class periods to complete the unit and the exercise sheets. Frequent discussion periods were held during which time the control-group students were limited to economic geography.

For the experimental group the above technique was employed and, in addition, use was made of the incidental method of instruction in an endeavor to influence pupils to become more internationally minded. The method consisted of carefully planned procedures to develop in the

pupils a feeling of intimacy for people of distant lands. It was an apparently spontaneous technique that called for interesting sidelights on the people and products studied. During the study of a region the teacher frequently turned from the direct consideration of the subject to mention a custom in that region, or to outline a problem of the people, to show that their problems are similar to the problems which confront us. Again, frequent mention was made of the achievements of other people, their heroes, and the traditions and ideals that they hold.

In order that there should be some definite directions towards which influencing might take place, it was decided to direct teaching towards increased respect for the Germans, increased opposition towards war, and an increased preference for the Chinese. For this purpose the three relevant Thurstone attitude scales were used. Form A of a scale was given to each group; then, after a period of four weeks during which time incidental instruction was given to the experimental group, Form B of the scale was administered. Also, at the end of the whole eighteen-week period of experimentation, the Neumann-Kulp-Davidson test was repeated. The following table shows the results. I.S. stands for initial score and E.S. for end score. The positive sign with the differences between means of gains indicates that the advantage was on the side of the experimental group. The Neumann-Kulp-Davidson test is scored in such a manner that low scores show cosmopolitanism and high scores provincialism. A similar thing is true of the war attitudes scale used. This must be kept in mind in interpreting the "advantage" in the table below.

TABLE V  
SUMMARY OF MEAN SCORES AND MEAN GAINS ON FOUR CRITERIA OF  
INTERNATIONAL-MINDEDNESS

Test	Control Group			Experimental Group			Differ- ence	Ratio to S.E.
	Mean I.S.	Mean E.S.	Mean Gain	Mean I.S.	Mean E.S.	Mean Gain		
Germans.....	6.990	6.997	.0075	6.930	6.975	.045	.0375	.155
War.....	4.692	5.052	.360	4.652	4.535	-.1175	.477	1.66
Chinese.....	5.437	5.975	.538	5.540	6.037	.497	-.041	.10
Neumann, Kulp, and Davidson..	3.979	4.028	.049	3.971	3.890	-.081	.130	1.60

Inspection of the table reveals that three of the tests show a greater growth of internationalism on the part of

the experimental group than on the part of the control group. With respect to the Chinese there was a slight difference in favor of the control group, but this difference was only one tenth of its standard error.

## II. EFFECTIVENESS OF THE OPAQUE PROJECTOR (STOVER)

One comes to know any country or people by living the life of that country or people. Since this is impossible of achievement to any considerable extent for many individuals and groups, we must depend upon vicarious participation and indirect observation through the medium of pictures, stories, etc. In the experiment about to be described, we employed such concrete aids in order to introduce our pupils realistically to the peoples of the world. We stressed in the races studied: (1) similar culture traits, (2) kindness and congeniality traits, (3) dependability traits, (4) certain races as victims of persecution and oppression, (5) the noncompetitive achievement of outstanding individuals and of the race in general. In addition, the classroom instruction stressed the effect of environment upon standards of living, the living conditions on various economic levels, explanation of causes of racial conflict, and examples of devotion of peoples to their chosen religion.

In the first experiment two groups of twenty-four ninth-grade girls each were paired on the basis of scores obtained from the Bogardus Racial Distance Scale. All of the girls were daughters of native-born white parents who had had little contact with racial groups other than their own. One of our groups received instruction in the form of eight illustrated travel talks with materials selected from the *National Geographic Magazine*, *Lands and Peoples*, and books dealing with the various races. The Negro was discussed mainly in the light of the achievement of prominent members of the race with pictures of Negro leaders available in *Who's Who in Colored America* and similar publications. The pictures were shown with an opaque projector and a translucent screen in a semidarkened room. The pupils were asked to make note of items about the



racess that served to change their opinion of the race in question for better or for worse.

The other group received as nearly as possible the same topics and descriptions of the same conditions of home life, etc., except that pictures were not shown. Tests were given again after twelve weeks. The results follow:

TABLE VI  
ATTITUDE CHANGES IN NINTH-GRADE GIRLS

Test	Mean Gain, Visual-Aid Group	Mean Gain, Non- Visual-Aid Group	Difference	S. E. of Difference	Chances True Difference Same Direction
Bogardus .....	-1.01	-.387	-.623	.197	1,300 to 1
Hinkley (Negro) ..	.316	.246	.07	.26	1.5 to 1
Neumann (International Attitudes) ..	-.208	-.262	.054	.11	2.2 to 1

On all three criteria the table shows appreciable gains by each group, both of which had received systematic instruction with the objective of making them more appreciative of races other than their own. These gains ranged from three to seven times their standard errors. The table also brings out the differential effect of the use of visual aids. The Bogardus Racial Distance Scale indicates a highly significant difference in gain in favor of the visual group, since the difference is more than three times its standard error. This difference is more clearly due to the controlled factors introduced into the experiment. The measurement of improvement of attitude towards the Negro shows a difference too small to be statistically significant, as is also true of measurement of growth in liberalism by the Neumann-Kulp-Davidson test. The former of these differences favors the visual group and the latter the control, since on both the Bogardus and the Neumann tests low scores lie in the direction of liberalism. These latter two tests lay somewhat aside from the main objectives of the experiment and were administered to measure certain possible concomitant liberal gains.

The experiment was repeated the second term with two sections of ninth-grade boys paired as in the preceding experiment. Due to matching difficulties (one small and one large section) and to absences, only fourteen pairs were

secured. But in this part of the study there was a certain modification of our procedure. Of twenty racial groups involved in the Bogardus scale (Armenians to Jews), the A section received instruction with pictures and the B section received the same instruction without pictures. On alternate days the B section received instruction on twenty groups (Jews to Welsh) and the A group the same instruction without pictures. In this experiment it was again found that both groups gained markedly in racial sympathies, the gain being more than ten times its standard error. It was also found, again, that both groups gained more on those themes on which visual aids were employed than on those on which they were not employed; the difference in favor of the visual aids was .96 times its standard error in the races A to J and 3.42 times its standard error in races J to W, involving chances of 5 to 1 in the former instance and 3,200 to 1 in the latter that the true difference is in favor of the visual aids.

An initial test on attitudes towards war by D. D. Droba (Form A) was compared with a final test (Form B) to investigate the assumption that liberal attitudes towards racial groups might have some effect upon their views of war as a method of solving international problems. The results for the two groups of boys involved in the second experiment show increasingly pacifistic attitudes by mean gains of seven per cent and eleven per cent of the mean scores of the initial test. Since both groups received visual aids, there is no opportunity to determine the relation of these to the gains.

Conclusions from the three experiments are:

1. The consistency with which our findings in these experiments point in the same direction amply confirms the thesis that international and interracial attitudes can be influenced by instruction governed by this objective, as far as the type of tests used in these experiments validly measure such development.
2. Visual aids seem to add appreciably to the effectiveness of education for international and interracial liberalism.
3. Gains in the function made the center of attention in the teaching are greater than those in the margin of attention, though some spread of liberalism in directions related to the central objective is indicated.

# THE TEACHING OF COURTESY IN THE JUNIOR HIGH SCHOOL

ALICE K. MILSOM

Because of lack of space only a single page can be allotted to this investigation. A fuller account will appear later in *The Pennsylvania School Journal*.

This study investigated the effect of the systematic teaching of ideals and techniques of courtesy in the junior high school of a Pennsylvania village. Courtesy was treated fundamentally as kindness; it was defined for the pupils by the nursery rhyme:

Politeness is to do and say  
The kindest thing in the kindest way.

Subjects were paired on scholastic marks. The teaching program on courtesy lasted three months. Initial and final measurements were taken in terms of ratings of pupils by one another. There was also a "delayed" measurement, three months after the close of the period of instruction. The following table summarizes the findings:

TABLE VII  
COMPARATIVE GAINS IN COURTESY RATINGS BY THREE GROUPS OF INSTRUCTED AND UNINSTRUCTED PUPILS

Grade	Number	Experimental			Control			S. E. of Difference
		Mean Initial Rating	Mean Final Rating	Mean Gain	Mean Initial Rating	Mean Final Rating	Mean Gain	
<i>Immediate</i>								
7th	20	3.28	3.30	.02	3.55	3.23	— .32	.34
8th	16	3.00	3.04	.04	3.27	3.28	.01	.03
9th	22	3.36	3.82	.46	3.67	3.49	— .18	.64
<i>Delayed</i>								
7th	20	3.28	3.34	.06	3.55	3.69	.14	— .08
8th	16	3.00	3.25	.25	3.27	3.39	.12	.13
9th	22	3.36	3.58	.22	3.67	3.44	— .23	.45

The table shows that the instructed group exceeded the uninstructed in gains in all three grades on the test at the close of the period of instruction and that these advantages were still prevailingly held at the period of delayed measurement, but by a somewhat reduced differential.

## WORKBOOK VERSUS ORAL INSTRUCTION

ELMER W. CRESSMAN

The purpose of the experiment was to determine whether or not character, or at least moral knowledge, could be improved by presenting to junior-high-school pupils life situations upon which they were called to pass judgment. It was further attempted to measure the relative gains made when the situations were presented in printed workbook form requiring written answers, against presenting the same situations orally by the teacher, the class responding in general discussion. Instruction by means of the workbook and oral presentation methods were in turn to be measured against the gains made by those having no direct moral instruction at all.

The work was carried on in the seventh grade of a large junior high school. The workbook selected was *What's the Right Thing to Do?* by W. W. Charters, Mabel F. Rice, and E. W. Beck, published by The Macmillan Company, 1931. This is the book assigned to the seventh grade in a series of workbooks called "Conduct Problems."

The selected workbook presents thirty-two well-chosen, lifelike conduct situations in a readable, interesting fashion. The pupil is confronted by a series of facts. Upon these he forms an opinion and makes a judgment. He determines for himself what he considers the right thing to do under the circumstances. The printed materials do not attempt to moralize. A series of printed questions calls the attention of the student to the various angles from which the problem may be viewed as well as giving the opportunity for a written reaction from every pupil. Only seventeen cases were presented, because this material became the subject for formal instruction in the forty-minute guidance period, one period per week for one term.

It was necessary to organize three matched sections of pupils. One group was to receive moral instruction by way of the workbook; the second group was to have the same cases presented orally by the teacher; and the third

was to serve as a control group. For purposes of matching and measuring gains, two tests and the I.Q.'s were used. Charters's workbook begins and ends with a summary statement of twenty situations. These he calls preview cases. To the twenty cases were added eleven more, so that each case to be used later for instruction would have at least one presentation in test-question form. When the preview cases were arranged in multiple-choice test form, they constituted a preview test which should have been useful in finding where the individuals stood in relation to the selected situations.

One sample of the added cases will give some idea of the nature of the test:

22. The law provides that children under 16 years of age may not operate a motor car. If you were 15 years of age and knew how to drive a car, check the statement which tells what you would do.

- ( ) Would not drive because it is against the law
- ( ) Would not drive because it is dangerous for children to drive
- ( ) Would take a chance in case of an emergency
- ( ) Would drive at any time because it is difficult to distinguish between a 15-year-old and a 16-year-old child

A standardized test of a more general nature was also desired. The Good Citizenship Test, developed in connection with the Character Education Inquiry and published by the Associated Press, claimed to test moral knowledge with a reliability of .835. The validity was not estimated.

The two above mentioned tests were administered to 320 seventh-grade pupils for whom I.Q.'s were available. These three elements constituted the basis for matching. In the Preview Test, each case scored one with a total possible score of 31. The Good Citizenship Test is made up of fifty elements each of which scores two with a total of one hundred. For purposes of matching, it was desirable to have the three scores combined into a single score. It would have been impossible to get forty-seven sets of triplets, identical in all three scores. The essential feature of any scheme of combining scores is that the variabilities of the component sets of scores shall be an-



proximately equal, since the greater the variability of a set of scores the higher its weighting becomes in the combination. To make the variabilities equal, it is necessary to prepare the scores for averaging by multiplying the scores in all sets, except one, by some factor.

The standard deviations of our tests were: Good Citizenship, 10; I.Q.'s, 11.7; Preview Test, 2.9. When roughly compared, the S.D. of the Good Citizenship Test and that of the intelligence quotients are about equal while the Preview Test has an S.D. about one fourth as great. We wished to give the Good Citizenship Test double weight because we considered it to be the most promising of the measuring elements. A learning score was, therefore, obtained for each pupil by summing four times his Preview Test score, twice his Good Citizenship Test score, and once his intelligence-quotient score.

From the 320 pupils tested, 141 were selected and matched into triplets on the basis of these composite learning scores, each set having the same average composite score. One of the triplets was assigned to the workbook group, another to the oral-instruction group, and the third to the control group. At the close of the experiment, 111 pupils, or 37 sets of triplets, remained in the experiment.

The authors of the Good Citizenship Test report a correlation of it with I.Q.'s,  $r$  equal to .614. The 320 cases used in this experiment gave this  $r$  equal to .38. The Preview Test correlates with the I.Q. .09.

Section A worked not more than one period each week upon each case in the workbook. The teacher distributed the work sheets and the pupils responded in writing without comment. It was necessary from time to time to give individual assistance with the reading of the materials. Section B was more interesting to watch. The teacher presented the conduct situation to the pupils in as stimulating a way as possible. The students responded with lively discussions as to what they would have done had they been confronted by the same conditions. Sometimes the arguments became heated. Following the lead of the ques-

tions printed in the workbook, the teacher led the discussion along the various angles. The discussion was never prolonged or forced. If less than a period was necessary to complete the lesson with either section, the work was allowed to end naturally.

When the seventeen lessons had been completed, the students remaining in the experiment were again given the Preview and Good Citizenship Tests in order that the individual and class gains might be measured. The Preview Test, being based upon the material used for purposes of instruction, was expected to test how well the pupils learned their lessons. The following summary table shows the mean scores for each group.

TABLE VIII  
COMPARATIVE GAINS BY THREE GROUPS ON TWO CRITERIA

	1 Workbook	2 Oral	3 Control	Differ- ence 1-3	Differ- ence 2-3	Differ- ence 1-2
<b>A. Preview Test</b>						
Mean initial score.....	20.08	17.64	19.06			
Mean final score.....	21.57	19.03	19.62			
Mean gain.....	1.49	1.66	.56	.93	1.10	-.17
S.E. of difference.....				.77	.72	.62
Chances true difference is in same direction.....				7 to 1	14 to 1	1.5 to 1
<b>B. Good Citizenship Test</b>						
Mean initial score.....	55.45	60.45	55.35			
Mean final score.....	68.00	63.35	59.70			
Mean gain.....	12.55	2.90	4.35	8.20	-1.40	9.65
S.E. of difference.....				1.84	2.3	1.47
Chances true difference is in same direction.....				"certain"	3 to 1	"certain"

It will be observed that on the Preview Test both instructed groups exceeded the uninstructed in mean gains, and the workbook group exceeded the uninstructed in the Good Citizenship Test, but on this latter test the oral group fell a little below the control. On the Preview Test the workbook group dropped a mere trifle below the oral group while on the Good Citizenship Test the workbook exceeded the oral by a large margin, the difference being 6.56 times its standard error. While the findings are far from conclusive, the prevailing directions and the relative sizes of the differences suggest that instruction on moral problems contributes somewhat to the clarification of the moral concepts of junior-high-school pupils, and that the workbook method seems superior to the oral, particularly in getting transfers to materials different from those used in training.

## INDIVIDUALIZED METHOD AND CHARACTER EDUCATION

GRACE E. ALLEN

This investigation sought to discover the comparative effectiveness of the teaching of plane geometry by the individual and the recitation methods of instruction in actual subject-matter achievement and in the development of certain personality traits. It is more clearly defined by resolution into the following questions: What is the effect of differences in teaching method upon student ability and upon traits of character in a given academic subject? If two groups are taught by the same instructor, and equated for initial ability, one group being taught by the method of individual instruction and the other by the traditional classroom method, what differences in subject-matter achievement are apparent at the end of the course? What is the degree and direction of change in the two groups in these personality traits: neurotic tendency, introversion-extroversion, dominance-submission, self-sufficiency, honesty, prejudice, and mathematical interest?

The experiment was conducted with two groups of eleventh-grade students in the Senior High School of Altoona, Pennsylvania, during the entire school term of 1932-1933. Each group contained approximately seventy-five students, who were divided into smaller sections for the purpose of group meetings. These sections of each group, however, were treated in the same manner. No student had studied the subject before the beginning of this study and all were of average or superior intelligence.

In order to adapt these pupils to the purposes of this investigation one group was subjected to the study of plane geometry under the traditional classroom method, the other to the method of individual instruction. The two groups were equivalent in subject-matter prerequisites, used the same textbook, covered the same amount of subject matter, met in the same classroom, were measured by the same tests, guided by the same instructor, and every effort

was made to keep all factors constant with the exception of the experimental factor—the teaching method.

By the traditional classroom method is meant the procedure in which the class period is divided into several recommended sections: first, a *review* of the previous day's work under the direction of the teacher and usually taking the form of a test or drill, oral or written; second, the *recitation* by the students on the material that had been assigned them the previous day for outside preparation, consisting chiefly of board proofs, criticisms and discussion, questions and answers; third, the *advance* lesson, in which the group, with the instructor's guidance, is led to develop a relationship between the present discussion of the subject and a new hypothesis, thus leading into the fourth section or the *assignment* of the lesson to be prepared for the following day. Time remaining is given to *supervised study*.

The method of individual instruction implies the placement of the responsibility for learning on the individual. This technique demands self-instructive and self-corrective practice for each student in order that he may study each unit of subject matter with a minimum amount of help from his teacher and associates. For this purpose each student is supplied with mimeographed instruction sheets covering each unit of work. These were composed by the instructor in accordance with the text and six standardized unit tests were used. The student was permitted to meet the requirements set forth in these sheets at his own rate, with the exception that a time limit was set for each unit of work in order to ensure adequate completion of the course. When a section of work was completed to the satisfaction of the student, he was required to pass an objective test over the material included. Failure to pass this test prohibited the student's going forward until remedial practice corrected his errors and made it possible for him to pass an equivalent test. The classroom was a laboratory. The students enjoyed "freedom in work." The instructor was accessible for conference and guidance at all times. A class demonstration or discussion was resorted to only when desired by the group.

A recent textbook, *Modern Course in Plane Geometry*, by Strader and Rhoads, provided the core of subject-matter requirement. The Lane-Green Unit Achievement Tests in Plane Geometry were used to measure subject-matter achievement by units. Two equivalent forms of this test were available, the second form being used when a retest was required. The 1932 form of the Coöperative Plane Geometry Test was used to measure final achievement. Ability in geometry was measured by the Rogers Test for Mathematical Ability—geometry section. The intelligence test used was the Terman Group Test of Mental Ability.

The number and kind of personality traits considered in this study was limited, largely because of the paucity of tests for such measures. Some personality traits which undoubtedly are affected by the individual method of instruction could not be measured because of the lack of testing materials. In some cases, however, there were several tests of a particular trait from which to choose and in those instances consideration was given to these features of the tests: usefulness, reliability, ease of administering, objectivity in scoring, validity, content for inclusiveness, and authorship.

The Bernreuter Personality Inventory measures several aspects of personality: neurotic tendency, self-sufficiency, dominance-submission, and introversion-extroversion. The reliability of the test is .86 and the validity .84.

The Self-Marking Test by Julius B. Maller measures the amount of deception an individual will express when opportunity for deception is given. The reliability of this test as given by the author is, by the Spearman-Brown formula, .92.

The Strong Vocational Interest Blank measures interest in many vocations. The measure of mathematical interest was applied in this study. In using the "odds versus evens" technique twelve coefficients of reliability for this test have been found which average approximately .80.

The Watson Test of Public Opinion measures objectively the tendency of any individual to manifest prejudice



and to measure the amount of that deviation from fair-mindedness. Its reliability is given as .96 and its correlation with criteria of validity as .85.

Pairing and matching were done separately for each personality trait tested and the different measures of achievement. Since there were approximately seventy-five students in each group, it was possible to match up fifty pairs for each measure under consideration with initial score disparities between members of the same pair so small, comparatively, as to be negligible. The bases for pairing were: (1) to measure achievement, initial scores of the achievement test used, intelligence quotients, and initial scores on the test of mathematical ability; (2) to measure personality traits, the initial scores in the respective tests.

The following table is a summary showing the comparison of the mean gains for the control and experimental group in achievement and in all the personality traits tested. In the first column is listed the trait tested; in the second column, the mean gain in the control group; in the third column, the mean gain in the experimental group. The difference between the mean gains is found in the fourth column. In column five the standard error of the difference between the mean gains is given, followed in column six by the ratio of the difference between the mean gains to the standard error of the difference, thus providing in the last column the chances that the true difference is on the same side.

TABLE IX  
COMPARISON OF MEAN GAINS FOR THE CONTROL VERSUS THE EXPERIMENTAL GROUP IN ALL MEASURES

Trait Tested	Mean Gain Control Group	Mean Gain Experimental Group	Difference in Mean Gain	S. E. of Diff.	Ratio	Chances
1. Achievement <sup>1</sup> .....	37.48	44.60	7.12	2.10	3.39	2900 to 1
2. Achievement <sup>1</sup> .....	37.40	42.78	5.38	2.62	2.05	48.5 to 1
3. Achievement <sup>1</sup> .....	37.86	45.52	7.66	3.49	2.19	69.1 to 1
4. Ability in plane geometry....	8.24	10.58	2.34	.744	3.16	1225 to 1
5. Neurotic tendency.....	11.60	2.01	-7.58	9.06	.836	4 to 1
6. Introversion-extroversion....	6.6	5.2	-1.4	5.33	.26	1.5 to 1
7. Dominance-submission.....	21.7	12.78	-8.92	7.53	1.18	7.4 to 1
8. Self-sufficiency.....	11.24	12.92	1.68	7.11	.236	1.5 to 1
9. Honesty.....	-1.14	-1.92	-.78	.712	1.09	6.4 to 1
10. Prejudice.....	-8.1	-11.32	-3.22	6.49	.496	2.2 to 1
11. Interest.....	-71.5	-90.3	-18.8	22.77	.825	1 to 3.9

<sup>1</sup>Pairing based on initial scores in Achievement Test

<sup>2</sup>Pairing based on intelligence quotient

<sup>3</sup>Pairing based on initial scores on Test of Mathematical Ability

This study shows that the group that was taught by the individual method was definitely superior to the group taught by the traditional recitation method in academic achievement. It also shows that changes did take place in certain personality traits of the pupils between the beginning and end of the course. In the experimental group the changes in personality traits when compared with the control group took the direction in favor of less neurotic tendency (more emotional stability), less introversion (more extroversion), less dominance (more submission), more self-sufficiency, less deception (more honesty), less prejudice (more broad-mindedness), and less mathematical interest.

In only one case can it definitely be said that the change was undesirable with respect to the experimental group; that is, in the measure of mathematical interest. In the case of academic achievement the differences are large enough as compared with their standard errors to carry good statistical significance. In all the other cases the reliabilities are low considered individually, but the fact that they point so largely in the same direction adds greatly to their significance.

It is obvious that the experimental ratios are consistently smaller for the incidental learnings than for the actual subject-matter achievement. This result is consistent with psychological belief, according to which the amount of improvement in a capacity trained is probably never accompanied by an equal amount of improvement in other capacities, which varies according as these compare with the one specifically trained. However, the results seem to justify the attention of educators to the new method of instruction, not only as a means of obtaining better results in academic achievement, but also in producing desirable changes in personality traits.

## THE RESULTS OF THE INCIDENTAL METHOD OF INSTRUCTION IN CHARACTER EDUCATION

F. R. KNISS, E. K. ROBB, AND E. A. GLATFELTER

For the purpose of determining the results of the use of the incidental method of instruction in character education, controlled experiments were set up in three Pennsylvania senior and junior high schools in connection with various courses of study.

### I

At Madera a study was undertaken to determine whether character could be taught incidentally in the instruction of the tenth-grade course in history (Kniss). The experiment was started in October 1932 and extended until May 1933. The socio-economic status of the pupils was secured by the use of the Sims Score Card for Socio-Economic Status, and the mental age was determined by the use of the Otis Self-Administering Test of Mental Ability. Two sections of the tenth grade were selected for the experiment. The pupils were matched on the basis of mental age and socio-economic status. Eighteen matched pairs were available for the experiment. All character instruction was incidental and led directly from the study of tenth-grade history.

The results of the experiment were measured by the use of two tests: (1) Baker, Telling What I Do, and (2) a test devised by the instructor. Both tests set up certain life situations to which the pupil has three possible responses. The Baker test consists of eighty situations and the instructor's test of twenty. These tests were used at the beginning and at the end of the experiment. The same instructor was in charge of both sections.

The results of the experiment as secured from the tests used for the pupils included in the experiment favored the control group, as shown in the table on page 260.

TABLE X

STATISTICAL RESULTS OF THE DATA SECURED FROM THE USE OF TESTS

<i>Tests</i>	<i>Difference in Means</i>	<i>S. E. of Difference</i>	<i>Ratio of Difference to S. E. of Difference</i>	<i>Chances That True Difference is in Same Direction</i>
Baker test .....	-2.1	4.4	.47	2.1 to 1
Teacher test .....	-2.6	1.4	1.85	30.1 to 1

It is therefore concluded that, in so far as these groups are concerned, the incidental instruction had no beneficial effect upon the experimental group as measured by the tests used. No pupil or teacher ratings were made.

## II

A study was made of the value of incidental instruction for character building on the junior-high-school level at Bedford, Pennsylvania (Robb). A controlled experiment was set up in the seventh, eighth, and ninth grades. The pupils were matched upon the basis of their intelligence quotients as determined by the Otis Group Intelligence Scale, and their socio-economic status as determined by the Sims Score Card for Socio-Economic Status. One control and one experimental group were provided for each grade.

In the control groups the work in each subject proceeded according to the regularly prescribed courses of study. In the experimental group the work proceeded in much the same manner, with the exception that frequent reference was made whenever possible in the class procedure and discussion to something concerned with character. Every effort was made to stimulate this discussion extemporaneously so as to avoid giving the pupils in the experimental group the impression that a prepared program in moral education was in progress. Such traits as industry, courtesy, usefulness, obedience, service, loyalty, patriotism, truthfulness, sportsmanship, honesty, tolerance, world-mindedness, and citizenship were stressed in each class when an opportunity was presented.

As a means of measuring the results of the experiment, a series of tests was used, as well as ratings secured from teachers and pupils. Special permission was secured from D. C. Heath and Company for the reproduction and use

of the discrimination tests included in Fishback and Kirkpatrick—*Conduct Problems for Junior High School Grades*. The Good Citizenship Test developed in connection with the Character Education Inquiry was also used for measuring the moral knowledge and ethical discrimination of the pupils. The pupil ratings in the junior high school were made by the use of the Character Education Inquiry Guess Who Test, and the teacher ratings by the use of the Character Education Inquiry Conduct Record Sheet. All of the tests and the pupil and teacher ratings were used in both control and experimental groups at the beginning and at the end of the experiment.

The results of the Bedford experiment are shown in the following table:

TABLE XI

COMPARATIVE ATTAINMENTS OF THE EXPERIMENTAL AND CONTROL GROUPS ON NINE CRITERIA

Tests	Grade	Difference in Means	S. E. of Difference	Chances to 1 That the True Difference is on the Side of:	
				Control Group	Experi- mental Group
Discrimination I .....	7	— .85	.943	4.4	....
Discrimination II .....	7	— .15	.714	1.4	....
Discrimination III .....	7	—1.85	1.019	26.8	....
Discrimination IB .....	7	— .65	.6	6.3	....
Discrimination IIB ...	7	.05	.836	....	1.12
Good citizenship .....	7	— .70	3.31	1.4	....
Discrimination I .....	8	.55	.728	....	3.4
Discrimination II .....	8	1.60	.889	....	26.8
Discrimination III .....	8	— .10	.616	1.32	....
Discrimination IB .....	8	.40	.574	....	3.1
Discrimination IIB ...	8	1.40	.793	....	24.0
Good citizenship .....	8	5.15	2.06	....	160.0
Discrimination I .....	9	.25	.331	....	3.4
Discrimination II .....	9	— .58	.583	5.2	....
Discrimination III .....	9	— .54	.556	5.0	....
Discrimination IB .....	9	— .39	.436	4.4	....
Discrimination IIB ...	9	— .75	.721	5.7	....
Good citizenship .....	9	— .11	2.21	1.1	....
Teacher rating .....	7	1.50	1.49	....	5.3
Teacher rating .....	8	3.05	2.15	....	11.5
Teacher rating .....	9	.57	1.59	....	1.8
Pupil rating .....	7	.15	.244	....	2.6
Pupil rating .....	8	.15	.3	....	2.2
Pupil rating .....	9	.05	.244	....	1.4



From the table on page 261 we find that the differences in the tests slightly favored the control groups in the seventh and ninth grades. In the eighth grade the differences as measured by the tests used favored the experimental group. The teacher and pupil rating in all grades of the experiment favored the experimental groups.

Since the value of character education is in its expression in the actions of the individual, the ratings on conduct are more valid measures than those on information or judgment. It is planned to remeasure the results of this experiment at the end of a year after the experiment is completed.

In the scoring of the tests used for measuring the moral knowledge and ethical discrimination it was found that the pupils already had a very acceptable amount of moral knowledge at the beginning of the experiment, which may have affected the results as far as these tests were concerned.

### III

The York experiment (Glatfelter) is still in progress, so that only preliminary findings are reported in this article. It involves nearly five hundred pupils in grades seven, eight, and nine of the Hannah Penn Junior High School. The pupils were matched for experimental and control sections on the basis of intelligence quotients, since these are known to correlate reasonably highly with desirable moral traits. The experimental factor consisted of incidental moral instruction similar to that described for the two preceding experiments. Attainment was measured by change in average ratings by pupils, and in average ratings by teachers, between the beginning and the middle of the year, and again between the middle and the end of the year. Forty-three teachers and five hundred pupils contributed towards the ratings. The ratings were secured on five character traits: coöperation, courtesy, industry, loyalty, and dependability. Ratings were taken on these traits one at a time, each on a separate day, and each trait was carefully defined for the raters.

So far the results have been worked up only for differences between first and second ratings. When the experiment is completed changes will be measured to at least a third and a fourth period. Of eight pairs of groups (boys and girls considered separately for grades 7B, 7A, 8B, and 9B rated by both teachers and pupils), the differences favored the experimental groups in seven and the control groups in nine in the trait of coöperation. In courtesy, eight differences favored the experimental and eight the control. In industry, nine favored the experimental groups and seven the control. In loyalty, nine favored the experimental and seven the control groups, and in dependability four favored the experimental groups and twelve the control groups. There is, therefore, nearly an equal division of the honors between the experimental and the control groups. This is not due to unreliability of the measuring instruments since, as indirect evidence of reliability, the scores on the several traits intercorrelated, though taken on different days, from .86 to .90 in the teacher ratings and from .89 to .93 in the pupil ratings. That it cannot be charged to lack of validity of the measuring instrument is suggested by the fact that the averages of the pupils' ratings for individual students correlated with the averages of teachers' ratings from .876 to .906 when corrected for attenuation. The failure to secure differential advantages for the instructed groups seems chargeable only to lack of functioning value in the experimental factor.

From this trio of experiments it seems clear that incidental instruction in morals is ineffectual in improving moral judgment and in furthering moral conduct.

## THE EFFECT OF ATHLETICS ON CERTAIN CHARACTER STUDIES

J. L. HACKENBERG, E. B. YEICH, AND L. A. WEISENFLUH

It has been a debatable question in the minds of many administrators as to whether athletics, as conducted in most secondary schools, do really contribute anything worth while to scholarship and character traits. A number of experiments have been made to see how athletic activities are related to scholarship, but very little has been done to see whether they contribute anything to character traits.

Three controlled experiments have been conducted during the past year, by the authors of this article, to get experimental evidence on this matter. These experiments are along the same general line, but differ in details. So we shall give a brief account of each experiment separately and then draw our conclusions from a composite of all three.

The first of these was conducted in the high school of Sandy Township, DuBois, Pennsylvania, by Mr. Hackenberg. The object was to ascertain whether organized athletics, as conducted in that school system, really contributed anything to certain character traits. The study of progress continued during the entire school year. The main sports in this school are football, basketball, and track.

In our school the student body may be divided into three groups or classes: those pupils who take active part in athletic contests between our school and other schools; those pupils who have no active participation in athletics but are interested in the sports, attend all games and all kinds of athletic meetings; and those pupils who do not participate in any athletics or do not attend any meetings of any kind; in fact, they are rather antagonistic to athletics.

We set up two parallel group experiments. We shall name the group that took active part, Group I; the group that took no active part, but was interested, Group IA; and the group that neither took an active part nor was interested, Group IB. The first experiment compared

Group I with Group IA and the second Group I with Group IB. We used forty cases in each group.

Members of the groups were paired on the following bases: mental age, achievement scores of the previous year, initial status of the pupils, curriculum followed in high school, sex, grades in school, and location in district. Only pupils of grades nine and ten were used in the study.

Six different tests, taken from the Character Education Inquiry battery of tests were employed to measure comparative growth: the Good Citizenship Test, the Information Test, information part of Self-Scoring Intelligence and Achievement Tests, O'Reilly's Character Analysis Chart, and the New York Rating Scale for School Habits.

We attempted to measure the following character traits: honesty, which we measured in the light of testing for truthfulness, whether the pupil is willing to accept deserved blame or whether he tries to lay the blame on some one else; citizenship, which we measured in the light of the pupil's ability to adapt himself into society; obedience, which we measured in the light of the pupil's ability and willingness to abide by the regulations of society; sportsmanship, which we measured in the light of the pupil's willingness to play fair in all things. Furthermore, we wished to find out whether athletics would help the participant to make worthy use of his leisure time.

We administered the three Character Education Inquiry tests to the entire school at the beginning of the term. We had the pupils rate themselves on the O'Reilly Character Analysis Chart and had the teachers rate the pupils in their respective homerooms on the New York Rating Scale for School Habits during the first week of school. These results were tabulated and recorded in the office of the superintendent. From then on the program was entirely forgotten, as far as the teachers and pupils were concerned, until almost the end of the school term. Then the same tests, or different forms of the same tests, were again administered under the same conditions as the initial tests, and again pupil and teacher ratings were made. These

results were then tabulated and compared with the initial scores and ratings. We then proceeded to work up our results statistically. A summarized statement follows:

TABLE XII

COMPARATIVE GROWTH DURING ONE SCHOOL YEAR OF THREE GROUPS ON TWO CRITERIA

## A. GOOD CITIZENSHIP TEST

	<i>Initial Test</i>	<i>Final Test</i>	
	<i>Scores</i>	<i>Scores</i>	<i>Gain</i>
Group I: Averages.....	34.65	36.20	1.55
S.D. ....	5.06	5.03	
Group IA: Averages.....	34.37	34.97	.60
S.D. ....	4.02	5.47	
Group IB: Averages.....	34.37	35.27	.90
S.D. ....	4.07	6.15	
	<i>Group I over IA</i>	<i>Group I over IB</i>	
Difference between mean gains....	.95	.65	
S.E. of the difference.....	.6952	.2528	
Ratio of difference gain to its S.E..	1.36	2.54	
Chances of true difference in same direction .....	10 to 1	184 to 1	

## B. INFORMATION TEST RESULTS

	<i>Initial Test</i>	<i>Final Test</i>	
	<i>Scores</i>	<i>Scores</i>	<i>Gain</i>
Group I: Averages.....	142.17	144.62	2.45
S.D. ....	2.47	2.56	
Group IA: Averages.....	141.75	141.92	.17
S.D. ....	2.34	2.54	
Group IB: Averages.....	141.52	141.75	.23
S.D. ....	2.47	2.71	
	<i>Group I over IA</i>	<i>Group I over IB</i>	
Difference between mean gains....	2.28	2.22	
S.E. of difference.....	.5056	.4424	
Ratio of difference gain to its S.E..	4.51	5.0	
Chances of a true difference in same direction .....	308,500 to 1	3,488,000 to 1	

The Self-Scoring Intelligence and Achievement Test was used as a measure of the pupils' honesty. We find that in the initial test the experimental group had three cases where dishonesty was shown and in the final test two of these cases disappeared and only one remained. But in the control groups the same number of cases of dishonesty appeared in the final test as in the initial test.



On the self-ratings of the O'Reilly Character Analysis Chart the athletic group made a mean increase in score between initial and final rating of 1.7. Group IA made an increase of .46 and Group IB an increase of .08. Thus the athletic group excelled one of the nonathletic groups by 1.24 and the other by 1.62. The standard errors of these differences are, respectively, .503 and .472.

The teacher ratings on the New York Rating Scale for School Habits did not lend themselves to a quantitative evaluation comparable with the other tests. Of Group I, fourteen members increased their rating within the experimental period, five decreased their ratings, and 21 remained unchanged; of Group IA eight increased in rating, five decreased, and 27 remained the same; while of Group IB nine increased their ratings, six decreased theirs, and 25 remained unchanged. Thus in both types of ratings the athletic groups improved slightly more than either of the nonathletic groups.

In the West Reading Experiment (Yeich) twenty athletes were matched with as many nonathletes in respect to sex, grade, and intelligence, an athlete being defined as "a member of an athletic squad who participates in all practices and is present as a probable or actual participant at all games of his chosen sport." Scores for four character traits were obtained from teacher ratings. In three of these traits the mean of the athletic group exceeded that of the nonathletic, as shown in the following table:

TABLE XIII

MEAN RATINGS OF ATHLETES AND MATCHED NONATHLETES IN FOUR CHARACTER TRAITS AT WEST READING HIGH SCHOOL

	Fellowship		Followership		Obedience		Honesty	
	Athletes	Non-athletes	Athletes	Non-athletes	Athletes	Non-athletes	Athletes	Non-athletes
Mean.....	2.35	2.22	2.39	2.34	2.34	2.50	2.58	2.52
Difference.....	.09		.05			.16	.06	
S.E. diff.....	.11		.10			.12	.09	
Ratio.....	.8		.5			1.33	.07	
Chances.....	4 to 1		2.3 to 1			10 to 1	1.1 to 1	

As a guide to the teachers in ratings, the four traits involved in the study were defined as follows:

1. Fellowship—recognizes and extols the good qualities of others and is tactful and kind regarding the faults of others

2. Followership—sacrifices self for the sake of the task and co-operates cheerfully for the good of the group

3. Obedience—abides by the regulations of the school and recognizes authority whether teachers or pupils are in charge

4. Honesty—plays fair and accepts deserved blame

The Old Forge High School experiment (Weisenfluh) was conducted in the same manner as that of West Reading. Fourteen pairs of pupils were involved. The athletes were found to exceed the nonathletes in only one of the four character traits—honesty—while the nonathletes exceeded in the other three. But the ratios of the differences to their standard errors ranged from only .41 to .67. In none of the three studies were the differences between the two types of students in academic achievement found to be significant.

Thus, out of the eight possible comparisons with respect to contributions to character traits in the West Reading and the Old Forge experiments as rated by teachers, four were in favor of the athletic groups and four in favor of the nonathletic. As far as these two trials are concerned, therefore, we get no evidence that participation in athletics favors the development of these traits more than non-participation. But the Sandy Township experiment showed some net advantage to the athletic groups where certain objective tests were employed. And it is worthy of note that in this experiment changes during the year rather than status were considered and, since only pupils in grades nine and ten were used, development was caught at the beginning of the growth curve where changes, if there were any, would have the best opportunity to show themselves. In Sandy Township it is the practice of the coach to make the development of character a deliberate objective of his training, as it is also to some degree at West Reading. So that, all in all, this trio of experiments suggests the mere possibility that athletics may be made to contribute slightly to the development of character traits. But it also suggests that the contribution is much smaller than it is often alleged to be.

## SUMMARY OF THE PENN STATE EXPERIMENTS ON THE INFLUENCE OF INSTRUCTION IN CHARACTER EDUCATION

CHARLES C. PETERS

In the series of experiments described in this issue of THE JOURNAL, 180 measured comparisons of experimental and control groups were made. But these were in terms of very different measurements with quite unlike units, so that they are not readily comparable. In order to bring them together into a single form so that we may draw inferences from the whole set, we shall reduce all differences to terms of "standard scores" by dividing each difference between means by the standard deviation of the two paired arrays combined. That will put all differences in terms of a single unit, called  $z$ . Twenty of our 180 experimental contrasts either had to do with effects on scholarship or were of a sort not reducible to  $z$  scores, so that we shall not include them in this summary. Eighty additional ones are from Mr. Glatfelter's experiment which is now only partially completed. For the sake of economizing space we shall merely indicate the distribution of these as to sign. They confirm the evidence given by the other twenty-six relating to incidental instruction in showing that such instruction is ineffectual in measurably modifying conduct. The other eighty contrasts we list in the summary table below, grouping them under headings according to whether the instruction was systematic and centered on a specific theme, whether it was incidental, or whether the conduct outcomes accrued collaterally from academic courses or other activities. The plus values (indicated by the absence of a sign) mean that the advantage favored the moral instruction while the negative signs mean that the advantage lay on the opposite side. Consistently signed differences under a section show for the set a highly reliable advantage in the direction indicated;

inconsistently signed differences (that is, those with nearly an even number in each direction) suggest little or no true advantage.

TABLE XIV  
DIFFERENCES IN STANDARD UNITS BETWEEN MEAN MEASURES OF GROWTH IN CHARACTER OF  
EXPERIMENTAL AND CONTROL GROUPS

Experimenter	Grade Level	Theme	Nature of Measure	Number of Pairs	$M_e - M_c$ $\sigma$
I. SYSTEMATIC INSTRUCTION					
Milsom	7	courtesy	pupil ratings	20	.47
Milsom	8	courtesy	pupil ratings	16	.04
Milsom	9	courtesy	pupil ratings	22	1.12
Eichler	9	social leadership	pupil ratings	20	.15
Eichler	12	social leadership	pupil ratings	22	.14
Merrill	10	social leadership	pupil ratings	36	.33
Robb	12	philosophy of life	verbal	26	.40
Robb	12	philosophy of life	pupil ratings	26	.05
Robb	12	philosophy of life	teacher ratings	26	.63
Faust	9	moral problems	self-rating	26	— .24
Faust	9	moral problems	verbal	26	— .16
Faust	9	moral problems	verbal	26	.15
Faust	9	moral problems	verbal	26	.51
Faust	9	moral problems	self-rating	26	.41
Campbell	10-12	attitude—Germans	Thurstone scale	40	.03
Campbell	10-12	attitude—war	Thurstone scale	40	.34
Campbell	10-12	attitude—Chinese	Thurstone scale	40	— .02
Campbell	10-12	international attitude	verbal test	40	.33
Stover	9	international attitude	verbal test	24	.58
Stover	9	international attitude	verbal test	24	.85
Stover	9	attitude—Negro	Thurstone scale	24	.46
Stover	9	attitude—Negro	Thurstone scale	24	.21
Stover	9	racial attitudes	Bogardus scale	24	1.48
Stover	9	racial attitudes	Bogardus scale	24	.55
Stover	9	racial attitudes	Bogardus scale	28	.75
Stover	9	attitude—war	Thurstone scale	28	.64
Cressman	7	moral problems	verbal	37	.34
Cressman	7	moral problems	verbal	37	.30
Cressman	7	moral problems	verbal	37	.68
Cressman	7	moral problems	verbal	37	— .13
II. INCIDENTAL INSTRUCTION					
Kniss	10	morality	tell what I do	18	— .28
Kniss	10	morality	verbal	18	— .87
Robb	7	morality	verbal	20	— .30
Robb	7	morality	verbal	20	— .06
Robb	7	morality	verbal	20	— .65
Robb	7	morality	verbal	20	— .33
Robb	7	morality	verbal	20	.02
Robb	7	morality	verbal	20	— .07
Robb	7	morality	teacher ratings	20	.29
Robb	7	morality	guess who	20	.17
Robb	8	morality	verbal	20	.23
Robb	8	morality	verbal	20	.49
Robb	8	morality	verbal	20	— .06
Robb	8	morality	verbal	20	.20
Robb	8	morality	verbal	20	.52
Robb	8	morality	verbal	20	.63
Robb	8	morality	teacher ratings	20	.42
Robb	8	morality	guess who	20	.19
Robb	9	morality	verbal	28	.14
Robb	9	morality	verbal	28	— .26
Robb	9	morality	verbal	28	— .28
Robb	9	morality	verbal	28	— .23
Robb	9	morality	verbal	28	— .32
Robb	9	morality	verbal	28	— .01
Robb	9	morality	teacher ratings	28	.11
Robb	9	morality	guess who	28	.06
Glatfelter	7-9	morality	teacher and pupil ratings	80 trials, 37 positive and 43 negative	
III. COLLATERAL CONDUCT OUTCOMES					
1. Latin with character-training objectives					
Meek	10	citizenship	verbal	20	.47
Meek	10	antipathy to war	Thurstone scale	20	.59
Meek	10	international attitude	verbal	20	.87

## 2. Geometry by individualized method

Allen	11	emotional stability	verbal	50	.16
Allen	11	extroversion	verbal	50	.05
Allen	11	submission	verbal	50	.23
Allen	11	self-sufficiency	verbal	50	.05
Allen	11	honesty	verbal	50	.17
Allen	11	fair-mindedness	verbal	50	.09
Allen	11	interest in mathematics	verbal	50	— .15

## 3. Athletics

Hackenberg	9-10	citizenship	verbal	40	.29
Hackenberg	9-10	citizenship	verbal	40	.22
Hackenberg	9-10	moral judgment	verbal	40	.62
Hackenberg	9-10	moral judgment	verbal	40	.61
Hackenberg	9-10	morality	self-rating	40	.56
Hackenberg	9-10	morality	self-rating	40	.71
Yeich	9-12	fellowship	teacher rating	21	.27
Yeich	9-12	fellowship	teacher rating	21	.17
Yeich	9-12	obedience	teacher rating	21	— .47
Yeich	9-12	honesty	teacher rating	21	.19
Weisenfluh	9-12	fellowship	teacher rating	14	— .12
Weisenfluh	9-12	fellowship	teacher rating	14	— .12
Weisenfluh	9-12	obedience	teacher rating	14	— .02
Weisenfluh	9-12	honesty	teacher rating	14	.21

An inspection of the table shows that 26 of the 30 differences under systematic instruction favored the experimental groups. The consistency with which these differences point in the same direction indicates high reliability for the finding that systematic moral instruction can aid in the development of character. But under incidental instruction 56 of the differences favored the control groups and 50 the experimental, about an equal division; therefore it is indicated that incidental moral instruction is ineffectual in modifying the sort of conduct we attempted to measure. That athletics can make desirable contributions towards character development is indicated with a low reliability, and that character traits can be made to accrue as by-products from certain methods of teaching academic subjects is strongly indicated.

But the differences are small even when positive, much smaller than optimists are in the habit of believing. In those types of procedures that yielded prevalingly positive differences the median one is about .4 of a standard deviation. I have determined, on the basis of reasonable assumptions which space does not permit explaining here, that a difference of .40 shows that the experimental factor, present in the one group and absent from the other, constitutes roughly ten per cent of the factors making for change in the criterion; a difference of one sigma, about twenty-four per cent determination; of two sigmas, forty



per cent; and of three or four sigmas, practically complete determination of the criterion. So it is suggested by our set of experiments that the sort of systematic moral instruction we attempted enabled us to control some ten per cent of the factors making for growth in character of the type measured within the period of the experiment. If we add to instruction drives towards desirable character through various other school processes, combining all of these into an optimum team, it is possible that we might extend this percentage, after allowing for overlapping, to perhaps twenty per cent or a little more. The other eighty per cent may be determined by factors outside our control.

It is obvious that the instruction in these experiments involved "indoctrination." Although the instructors invited free discussion and challenge of every suggestion, it remains true that the teachers themselves believed that certain ways are "better"; that kindness, courtesy, peace-loving, etc., are better than their opposites—and the weight of the teacher's own convictions would inevitably count heavily in influencing the conclusions at which the discussions arrived. The resulting mass of ideas and convictions about right and wrong will be tested through all the future experience of the pupils in competition with counter ones, which will be from time to time suggested from other sources. If the insights and ideals to which the investigators helped their students are sociologically sound ones, it may reasonably be expected that they will grow and ultimately prevail; if they are "unfit," they will be overwhelmed and eliminated in competition with those suggested by other experiences.

## BOOK REVIEWS

*Child Psychology*, by BUFORD J. JOHNSON. Baltimore: Charles C. Thomas, 1932, 439 pages.

An excellent textbook on the psychology of early childhood following the more traditional organization of materials, but packed with fresh and interesting experimental data. An outstanding book in its field for courses attempting a systematic presentation of this developmental period.

*Development of Learning in Young Children*, by LOVISA C. WAGONER. New York: McGraw-Hill Book Company, Inc., 1933, 322 pages.

A study of the psychology of learning as it applies to the acquiring of early habits in the preschool child. An excellent discussion of the child's first experiences with the requirements of the adult world, and of his reaction to those requirements, a stage in which the fundamental character pattern is laid down. One of the McGraw-Hill Euthenics Series.

*Interrelations in the Behavior of Young Children*, by RUTH E. ARRINGTON. New York: Bureau of Publications, Teachers College, Columbia University, 1932, 156 pages.

An experiment in the development of reliable techniques for the facilitation of normative studies of behavior, but including significant data on the development of the child's attitudes towards his physical and social development. One of the Child Development Monographs and like all of them based upon admirable experimental work.

*Motivation of Young Children*, by LUCILE CHASE. Iowa City: University of Iowa, 1933, 119 pages.

An experimental study of the influence of various types of external incentives upon the performance of a task. Discussion of other similar studies. Interesting discussion of experimental technique. Bibliography on motivation.

*Youth and Sex*, by MEYRICK BOOTH. New York: William Morrow and Company, 1933, 299 pages.

The relationship of sex to the problems of adolescent adjustment, with particular emphasis upon the changing pattern of our social life. Emphasis upon the relationship of the question to educational procedures.

*Our Movie Made Children*, by H. J. FORMAN. New York: The Macmillan Company, 1933, 288 pages.

A summary report of the influence of the motion picture upon the attitudes, values, and character of youth, based upon the research in this field sponsored by the Payne Fund and the Motion Picture Research Council. Individual studies to be published by the Macmillan Company.

*Motion Pictures and Youth: A Summary*, by W. W. CHARTERS, and *Getting Ideas from the Movies*, by P. W. HOLADAY and G. D. STODDARD. New York: The Macmillan Company, 1933, 168 pages.

Combined within the covers of this volume are a summary statement of the purposes, methods, and scope of the Payne Fund studies, and the first report on the influence of movie experience upon the ideas of the child—an attempt to get at the ways in which and degree to which children learn from the movies.

*The Emotional Responses of Children to the Motion Picture Situation*, by W. S. DYSINGER, and *Motion Pictures and Standards of Morality*, by C. C. PETERS. New York: The Macmillan Company, 1933, 403 pages.

Combined within the covers of this volume are two more of the Payne Fund studies. The first is a rather technical study of children's emotional reactions to motion-picture situations. The second will arouse widespread interest, not merely because of its study of the influence of the movie on children's moral standards, but also because of the study of our moral standards (made by a most ingenious method) against which the movie influence is evaluated.

*Children—Why Do We Have Them?* by DORA RUSSELL. New York: Harper and Brothers, 1933, 287 pages.

A fresh, stimulating analysis of the parent-child relationship in the modern family, emphasizing the traditional attitudes that adults bring to family life, the satisfactions that they attempt to derive from it, the success that they have in deriving these satisfactions in the family of today, and the effect upon the child of this adult drama.

*The Family*, by KATHERINE D. LUMPKIN. Chapel Hill: University of North Carolina Press, 1933, 184 pages.

An unusually keen analysis of the social pattern of the modern American family. Particularly interesting is the analysis of the rôles the various members of the family adopt and the sources of the attitudes that determine those rôles. An excellent statement of the sociological approach to family life. One of the University of North Carolina Social Study Series.

*Parents, Children and Money*, by SIDONIE M. and BENJAMIN C. GRUENBERG. New York: Viking Press, 1933, 219 pages.

An interesting handbook for parents and teachers, dealing with the problems involved in introducing the child to the meaning and management of money. The first significant book in its field and an invaluable addition to any library on character education.

*Some Experiments in Living*, by PETER AINSLEE. New York: Association Press, 1933, 190 pages.

The author, a Protestant minister, has written an autobiography of selected experiences of his life. Although reminiscent in character, it presents an ardent plea for pacifism, international and interracial understanding, Christian unity, fidelity in marriage relationships, and a devout faith in God.

*Conflicts of Principle*, by ABBOTT LAWRENCE LOWELL. Massachusetts: Harvard University Press, 1932, 161 pages.

As the author states in the preface, "This little book contains no facts not perfectly well known to every one; and it avoids expression of opinion." It is an attempt to point out the proper limits between pairs of mutually contradictory principles. Through a series of short chapters, the author concisely reviews a series of conjugate principles in the fields of economics, politics, race, law, patriotism, education, personal conduct, mental patterns, and the teaching of Christ.

*Modern Tendencies in World Religions*, by CHARLES SAMUEL BRADEN. New York: The Macmillan Company, 1933, xi+343 pages.

To those who have assumed that religion is purely static, the data on actual changes in religion during the present century which are clearly presented by the author will come as a challenge. The major religions of the world, exclusive of Christianity, are presented in the light of their recent tendencies.

#### BOOKS RECEIVED

*Abstracts of Studies in Education at the Pennsylvania State College, Part III* (1933), Penn State Studies in Education, No. 8, edited by CHARLES C. PETERS and GEORGE W. HARTMANN. State College: School of Education, Pennsylvania State College.

*Adolescent Psychology*, by ADA HART ARLITT. New York: American Book Company.

*Behind the Doctor*, by LOGAN CLENDENING. New York: Alfred A. Knopf.

*Challenge of Humanism*, by LOUIS J. A. MERCIER. New York: Oxford University Press.

- Child Upbringing and the New Psychology*, by RICHARD AMARAL HOWDEN. New York: Oxford University Press.
- Christian Mission in the Modern World*, by WILLIAM D. SCHERMERHORN. New York: The Abingdon Press.
- Crisis of Democracy*, Annals of the American Academy of Political and Social Science, September 1933. Philadelphia: American Academy of Political and Social Science.
- Dawn of Conscience*, by JAMES H. BREASTED. New York: Charles Scribner's Sons.
- Dynamic Social Research*, by JOHN J. HADER and EDUARD C. LINDEMAN. New York: Harcourt, Brace and Company.
- Effect of Participation in Athletics Upon Scholarship Measured by Achievement Tests*, by JOHN ANDREW COOPER. Penn State Studies in Education, No. 7. State College: School of Education, Pennsylvania State College.
- Eugenic Predicament*, by S. J. HOLMES. New York: Harcourt, Brace and Company.
- From Peasant to Collective Farmer*, by N. BUCHWALD and R. BISHOP. New York: International Publishers.
- Institute for Child Guidance Report, 1927-1933*, by LAWSON G. LOWREY and GEDDES SMITH. New York: Division of Publications, Commonwealth Fund.
- Introduction to Western Civilization*, by GEORGE A. HEDGER, et al. Garden City: Doubleday, Doran and Company.
- It's Up to the Women*, by MRS. FRANKLIN D. ROOSEVELT. New York: Frederick A. Stokes Company.
- Man Into Woman*, edited by NIELS HOYER. New York: E. P. Dutton and Company.
- Management and Teaching Technique*, by GEORGE A. RETAN. New York: Prentice-Hall, Inc.
- Modern Tendencies in World Religion*, by CHARLES SAMUEL BRADEN. New York: The Macmillan Company.
- Molders of the American Mind*, by NORMAN WOELFEL. New York: Columbia University Press.
- Mother's Encyclopedia*, Four volumes compiled and edited by the editors of *The Parents' Magazine*. New York: Home and School Book Service, The Parents' Publishing Association, Inc.
- Newspaper Reference Methods*, by ROBERT W. DESMOND. Minneapolis: University of Minnesota Press.
- Social Behavior Patterns*, Volume I of "Observational Studies of Social Behavior," by DOROTHY S. THOMAS, ALICE M. LOOMIS, and RUTH E. ARRINGTON. New Haven: Institute of Human Relations, Yale University.
- Psychology and the New Education*, by S. L. PRESSEY. New York: Harper and Brothers.
- Sex in Childhood*, by ERNEST R. and GLADYS HOAGLAND GROVES. New York: The Macaulay Company.
- What We Are and Why*, by LAURENCE H. MAYERS and ARTHUR D. WELTON. New York: Sears Publishing Company, Inc.



